错觉

THE
AI
DELUSION

AI如何通过数据挖掘误导我们

[美]加里・史密斯（Gary Smith）著
钟欣奕 译

是智能
还是服从？

- 大数据、坏数据、假数据
  充斥着我们的生活
- 以相关关系取代了因果关系

中信出版集团

版权信息

书名:错觉：AI如何通过数据挖掘误导我们

作者:[美]加里·史密斯

译者:钟欣奕

ISBN:9787521709957

中信出版集团制作发行



第1章
智能还是服从

> 《危险边缘》是一档热门的电视智力竞赛节目，有多个版本，开播至今已有50多年。该节目的比赛内容为百科知识问答，其巧妙之处在于：参赛者要根据以答案形式提供的各种线索，给出与这个答案相对应的问题。例如，线索是“美国第16任总统”，正确的问题就是：“谁是亚伯拉罕·林肯？”每期节目均有三名参赛者，以摁按钮的方式口头抢答（除了最后一轮“终极危险边缘”以外，在其他环节三名参赛者均有30秒时间书写作答）。

从很多方面来看，这档节目都适合计算机参与，因为计算机能准确无误地存储和检索大量信息。在《危险边缘》青少年组比赛中，一名男孩因将“谁是安尼·弗兰克”误写成“谁是安妮·弗兰克”而痛失冠军。而计算机就不会犯这样的错误。

另外，线索有时通俗易懂，有时却晦涩难解。例如，线索是“把它打进去，你就输了比赛”，对只是资料库的计算机来说，很难得出以下正确问题：“什么是（台球）母球？”

还有一个难解的线索是：“翻译时，这支大联盟棒球队的名字会重复一次。”正确问题为：“什么是洛杉矶天使队？”（What is the Los Angeles angels？）

2005年，15名IBM（国际商用机器公司）的工程师合作设计了一款能与《危险边缘》最佳玩家同台对擂的计算机，取名“沃森”，以纪念IBM的首任CEO（首席执行官）托马斯·J.沃森。沃森在1914年接手IBM时，IBM还只是一家仅有1 300名员工、年收入不足500万美元的小公司，到了1956年他去世的时候，IBM已经发展成为一家有7.25万名员工、年收入9亿美元的公司。

“沃森”程序存储了相当于2亿页纸的内容，每秒可处理相当于100万本书的信息。除了拥有海量内存和高速处理能力外，“沃森”还能理解自然语言，使用合成语音进行交流。与罗列相关文档或网站的搜索引擎不同，“沃森”可按照程序并根据线索得出具体答案。

“沃森”运用数百个软件程序，先识别线索中的关键字和词组，再与海量数据库中的关键字和词组相匹配，最后得出合理答案。按照编好的程序，如果线索是某个名字（如亚伯拉罕·林肯），“沃森”就会写出以“谁是……”开头的问题；如果线索为某一事件，它就会写出以“什么是……”开头的问题。单个软件程序与某个答案的一致性越高，“沃森”就越能确定此为正确答案。

该程序能轻而易举地得出与“美国第16任总统”这么直白的线索对应的问题，但要处理有多重含义的词语时就有些困难了，比如，线索是“把它打进去，你就输了比赛”之类的问题。但是，“沃森”不会感到紧张，也绝不会遗忘。

2008年，“沃森”做好了参加《危险边缘》的准备，但还有些问题需要协商。IBM团队担心该节目的工作人员会使用包含双关语和具有双重含义的线索，给“沃森”下圈套。这一担心也恰好揭示了人类与计算机的巨大差异。人类可以根据语境理解词义，所以能理解双关语、笑话、谜语和讽刺批评。而目前的计算机，充其量只能检查出数据库中是否含有双关语、笑话、谜语或讽刺批评。

对此，节目工作人员同意随机抽取以往编写但未使用的线索。而节目工作人员也担心，如果“沃森”一得到答案就可以发出电子信号，会比必须通过摁按钮来答题的参赛者更有优势。对此，IBM团队同意给“沃森”装根电子手指来摁按钮，但它还是比人类快，这也让“沃森”占据决定性优势。摁按钮快算是聪明的体现吗？如果“沃森”的反应速度降为与人类的一致，比赛结果又会如何？

接下来，在2011年的人机大战中，“沃森”与《危险边缘》的两名前冠军肯·詹宁斯和布拉德·鲁特展开了两轮比赛。首轮比赛“终极危险边缘”的线索是：

它最大的机场以第二次世界大战的英雄命名，

它的第二大机场以第二次世界大战的战役命名。

两名前冠军给出的问题为：“芝加哥是什么？”而“沃森”给出的问题是：“多伦多是什么？？？？？”显然，“沃森”识别出了“最大的机场”、“第二次世界大战的英雄”和“第二次世界大战的战役”这些词组，然后在其数据库中查找相同主题，但没能理解线索的第二

部分（“它的第二大”）指的是该市的第二大机场。“沃森”给问题添加了多个问号，因为它计算出的这一答案的正确概率仅为14%。

尽管如此，“沃森”还是以77 147美元轻松获胜，詹宁斯和鲁特的赛果分别为24 000美元和21 600美元。“沃森”夺得了100万美元的冠军奖金（IBM将其捐赠给了慈善机构），詹宁斯和鲁特也各自将奖金的一半捐赠给了慈善机构。“沃森”在《危险边缘》的取胜是一次价值数百万美元的宣传良机。在获得艳惊四座的胜利后，IBM宣称，相比在《危险边缘》中与主持人亚历克斯·特雷贝克较量，“沃森”的问答技能将运用于更重要的领域。IBM一直将“沃森”应用于医疗、银行、技术支持以及其他能利用庞大的数据库来解决具体问题的领域。

对许多人来说，“沃森”击败《危险边缘》的两名前冠军无疑证明强大的“沃森”无所不知！计算机比人类更聪明，我们应该依靠它，相信它的决策。也许我们还应该担心，计算机会在不久的将来征服甚至消灭人类。

“沃森”真的比我们聪明吗？它的胜利恰恰反映了计算机的优势和弱点。作为能力超强的搜索引擎，“沃森”可以在其庞大的数据库中快速查找单词和短语（它还有可以快速点触的电子手指）。我之所以没有使用“解读”这个词，是因为“沃森”并不了解那些单词和短语的含义，比如“第二次世界大战”和“多伦多”，它也不明白语境中的词义，比如“它的第二大”。“沃森”的实力被过分夸大了，正如很多电脑程序一样，它的智能不过是假象罢了。

从很多方面来说，“沃森”的表现就是骗人的把戏，只不过是在范围极小的某些技能上看似具有超人的发挥罢了。设想有一个不懂英语，但有无限时间翻阅大型文库（藏有2亿页英语单词和短语）找出匹配单词和短语的人。我们会认为这个人聪明吗？计算机仅因能比人类更快地进行搜索匹配，就说明它聪明绝顶吗？

连IBM“沃森”团队负责人戴夫·费鲁奇也坦承：“我们在开发‘沃森’，设法让其仿造人类认知时，有坐下来好好谈过吗？根本没有。我们不过是想发明一台可以在《危险边缘》中获胜的机器而已。”

计算机不仅击败了《危险边缘》中的人类玩家，还击败了国际跳棋、国际象棋和围棋的世界冠军，这助长了人们认为计算机比最聪明的人类还要聪明的普遍观念。想要玩好这些战略型棋盘游戏，仅靠匹配单

词和短语的强大搜索引擎是远远不够的，还要能分析棋盘格局、制定创意策略、做到未雨绸缪。这难道不是真正的智能吗？

接下来，我们就从非常简单的儿童游戏开始了解。

井字游戏

在玩井字游戏时，两个玩家在3×3网格上轮流画×和 ○（如图1.1所示）。无论是在水平方向、垂直方向还是在对角线上，只要三个方格连成一条线，该玩家即赢得比赛。

通过分析所有可能的移动序列，软件工程师可以编写出靠蛮力计算的程序来掌握井字游戏。玩家甲有9个方格可选择，在他走出第一步后，玩家乙有8个方格可选择，前两步共有72种组合方式。走完前两步，玩家甲剩7个方格可选择。整局游戏玩下来，计算机程序必须处理的选择序列共有9×8×7×6×5×4×3×2×1 = 362 880种。

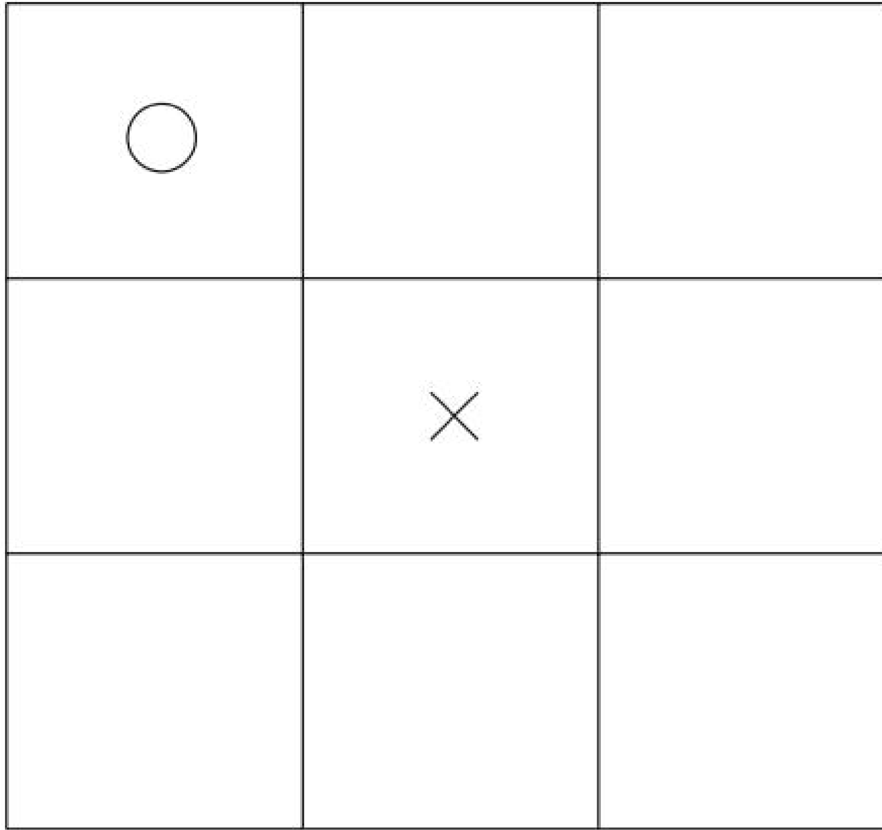


图1.1 井字游戏

也有更简便的分析方法，但重点是，井字游戏程序看待游戏的方式与人类不同。人类看到3×3的网格会思考选择哪些方块能完成三格连线，以及选择哪些方块会阻挡对手完成连线。但计算机程序无法对这些方格进行可视化，而是为每个方格分配一个1~9的数字（如图1.2所示），并识别获胜组合（例如1、2、3和1、5、9）。

1	2	3
4	5	6
7	8	9

图1.2 匹配数字后的井字游戏

计算机程序会算出1~9的可能序列，识别各玩家的最佳策略，并假设对手会选择的最佳策略。一旦软件编写调试完成，就会立即显示出最佳策略。

假设玩家乙采用最佳策略，如果玩家甲从中心格或任一边角格起步，玩家乙就选择相反的方式——如果玩家甲选择中心格，玩家乙则选择边角格；如果玩家甲选择边角格，玩家乙则选择中心格。采用最佳策略的游戏总会以平局结束。

这就是蛮力计算，不涉及逻辑推理，只是无意识地枚举数字1~9的排列和识别获胜排列。

在井字游戏和其他游戏中，人类通常会避免对所有可能的移动序列进行蛮力计算，因为这样一来移动序列的可能性就会暴增。相反，我们使用逻辑推理，并将注意力集中在有意义的走法上。与蛮力计算程序

不同，人类不会浪费时间思考明显错误的步骤。而没有逻辑和常识的计算机却还是会分析愚蠢的策略。

玩井字游戏时，人类玩家可能会研究3×3网格，而计算机玩的是1~9的数字。人类会采用可视化的方法，将注意力集中到中心格上，意识到这一格蕴含四个获胜排列，而每个边角格蕴含三个，每个边格蕴含两个。

中心格也是极佳的防守走法，因为接下来玩家乙无论选择哪一格，最多只蕴含两个获胜排列。相反，若玩家甲先选边角格或边格，就会让对方占据中心格，减少了自己的一个获胜排列，同时为对方创造了三个获胜机会。

从逻辑上讲，似乎起步最好选中心格，最后选边格。人类对棋盘的这种可视化认知和对中心格战略价值的判断，完全不同于软件程序对数字1~9所有排列的无意识识别。

人类也能发现游戏的对称性，即四个边角格任选其一开盘都同样可取（或不可取）。因此，人类只需思考选择其一的后果，选择其他三个边角格的后果就同理可得。游戏的对称性让人类每走一步都能减少需要考虑的移动步数。最后，人类会发现某些走法能迫使对手选择对其不利的方格，从而阻止对手完成三格连线。

人类能够运用战略性思维找出最佳策略，并发现采用最佳策略总会打成平局。有经验的人还会发现，同孩子玩游戏时不按常规出牌有时也能够获胜，例如开局选择边角格，甚至边格。

具有讽刺意味的是，尽管人类可以运用逻辑找出最佳策略，但人类编写的计算机软件程序还是有可能击败人类的，因为计算机无须考虑自己的走法。井字游戏的计算机程序只要遵守编程规则即可。相比之下，人类每走一步都必须思考，最后会疲惫不堪导致犯错。

计算机相比于人类的优势跟“智能”的一般含义毫无关联。正是人类编写出能识别最佳策略并存储于计算机内存中的软件，计算机才有规律可循。

尽管井字游戏这款儿童游戏会越玩越无聊，但它是很好的例子，凸显了计算机软件的威力和局限性。计算机程序对于烦琐的计算用处极

大，编程软件每次的答案都完全一致，还能不厌其烦地完成已编程好的任务。与人类相比，计算机的处理速度更快、保存的信息更多。

人类怎能奢望在以信息记忆和处理速度为胜的活动上与计算机竞争呢？也许真正的奇迹不是计算机的强大，而是人类还在很多方面比计算机更胜一筹。遵循规则与人类毕生所获得的智慧，两者天差地别。

人类的智慧使我们能够识别出含义模糊的语言和扭曲的图像，对问题追根溯源，应对异常情况以及很多虽遵循规则却无法处理的事情。

国际跳棋

国际跳棋比井字游戏复杂得多，实际上，它复杂到根本无法对所有可能的移动序列进行蛮力分析。所以，你可能认为计算机必须模仿人类的思维才能下好国际跳棋。然而并非如此。

国际跳棋采用灰白棋格相间的8×8棋盘（如图1.3所示）。只能走灰色棋格，也就是说可走方格的数量从64个减到32个。两方玩家每方各有12枚棋子，放置于己方的灰色棋格内，中间的8个灰色棋格留空。棋子可沿灰色棋格对角移动，跳吃对方棋子。

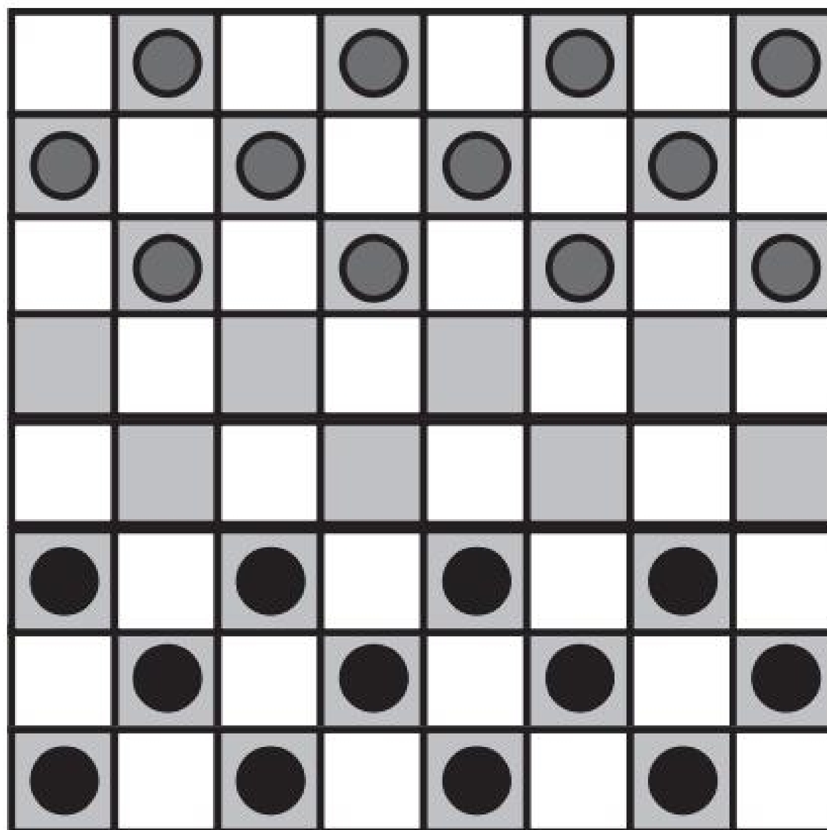


图1.3 国际跳棋棋盘

在理论上，尽管所有可能的序列都有无限步数，蛮力分析还是可以识别出最佳策略的，就像玩井字游戏那样。但是，对目前的计算机来说，在合理时间范围内要分析的可能序列数量过于庞大。因此，人类想出了简化策略来利用计算机的能力。与井字游戏一样，国际跳棋的计算机程序不会尝试制定逻辑策略，而是利用计算机的优势——快速处理和绝佳记忆。

井字游戏走九步就结束了，而国际跳棋有无限步数，因为玩家可以在没有哪方获胜的情况下不断来回移动棋子。实际上，来回移动棋子很无聊，所以除非有一方犯了非常低级的错误，否则玩家会在明显无法出现赢家时同意和棋。（冷酷无情的国际跳棋程序永远不会同意和棋，而是会一直玩到人类对手精疲力竭，累到无法清晰思考而犯错。）

虽然国际跳棋游戏的步数不受限制，但能走的棋盘位置还是固定的。用不着算出所有可能的移动序列，国际跳棋计算机程序更好的做法是查看所有可能的棋盘位置，然后确定在这些位置的走法哪些得势、哪些失势。

尽管如此，这项任务还是让人发怵。棋盘走位有5万亿种可能，在没有考虑到所有接下来可能的位置序列的情况下，很难真正确定走这一步是否会得势。

人类以其洞察力将游戏分为三部分（开盘、中场和残局），单独分析每个部分，最后串联起来（如图1.4所示）。

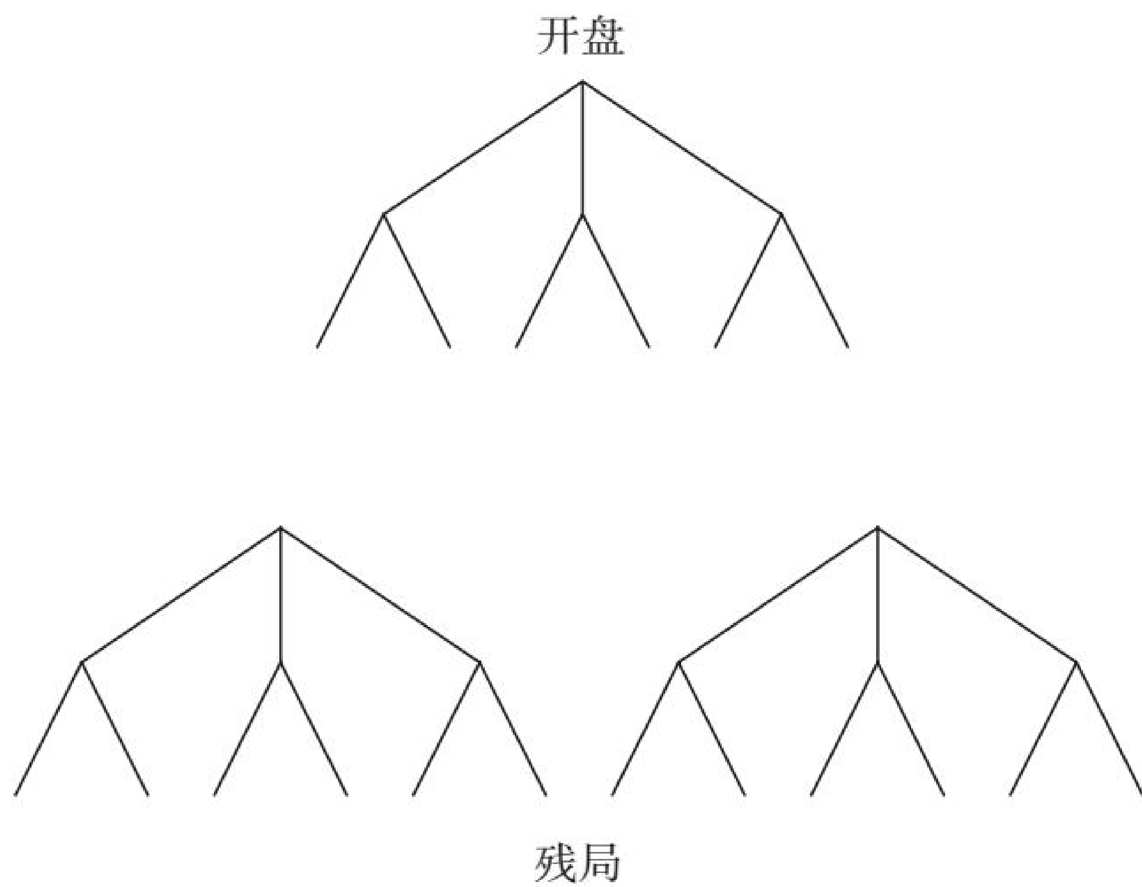


图1.4 国际跳棋的决策树模型

开盘的那几步棋已有写好的“剧本”，表明了最佳的开盘走法、每种开盘的最佳应对方式等。这些剧本是国际跳棋玩家几百年积累下来的集体智慧。每名严肃的国际跳棋玩家都会研究这些剧本。编写国际跳棋程序代码的软件工程师也会把剧本加载到计算机内存中，计算机会在开盘时遵守这些规则。

到了残局阶段，如果棋盘上只剩两枚棋子，则位置数量相对有限，而如果剩下三枚棋子，则位置数量会增加，但还在可控范围，以此类推。对于每个可能的位置，人类玩家能计算出最佳走法，同时确定最佳走法是否会造成平局或出现胜方。所剩棋子数越多，可能位置的数量就越多，但很多都容易解决，并且棋盘的对称性也会减少必须分析的位置的数量。人类分析完包含所有可能棋盘位置的全部残局的情况后，比方说还剩不到六枚棋子，那么每个位置的残局最佳走法就会被加载到计算机内存中。

游戏进行到预先加载的残局位置时，计算机便按照人类预先确定的最佳走法的规则落子。人机对抗的跳棋残局中，人类玩家每走一步，计算机就会从数据库中选出预先确定好的最佳走法来应对新的棋盘位置，一直持续到比赛结束，通常结果是一方认输或双方同意和棋。

在游戏进行到中场时，计算机会试图将开盘剧本与残局位置联系起来。如果开盘几步之后，游戏进行到已存储的残局位置，则游戏结果可想而知（假设为最佳玩法）。

可供蛮力分析来识别最佳序列的中场局势数不胜数，因此程序员会将人类在跳棋领域的智慧与计算机的能力结合起来，列举各种序列。如果计算机有足够的能力和时间预测接下来的四步，那么计算机就会预测这四步可能产生的所有序列，并使用人类特定的损失函数（loss function）来比较四步后所有的可能位置。损失函数也是基于人类几个世纪的经验，考虑了被认为重要的因素，例如，每个玩家拥有的棋子数和对棋盘中心位置的控制。国际跳棋专家建议程序员为不同因素分配权重，以反映每个因素的重要性。

计算机通常会选择“最大最小值归一化”（minmax）的走法，因此它可以在最坏情况下（即最大值）让可能造成的损失最小化（即最小值）。如果另一个玩家采用最佳走法，程序则选择损失最小（或收益最大）的走法。

经过几个回合的中场比拼后，棋子数缩小到前瞻计算可以得出已知残局结果的水平。假设这是最佳玩法，那么游戏基本上结局已定。如果人类玩家犯错，则游戏结束得更快。

值得注意的是，计算机程序中“智能”的含量极少。在游戏开始时，计算机程序必须遵守开盘提示；中场游戏期间，计算机程序确定前瞻序列，并使用人类规定的损失函数，按部就班地决定走法；进入最后阶段，计算机程序还得依照残局指令运行。

为国际跳棋、国际象棋、围棋等复杂游戏而设计的计算机程序并不试图模仿人类思维，这涉及对潜在取胜原则的创造性认识。编写计算机程序是为了利用计算机的优势——无懈可击的记忆能力和毫无差错的规则遵守。

国际跳棋的计算机程序与人类玩家相比有几个重要优势：它永远不会在开盘和结束时犯错。人类玩家可能已研究过国际跳棋手册，但人类没有完美的记忆能力，还是会犯错。没有人思考过，更不用说记住所有可能出现的残局序列，其中有些还需要几十步精确走法才能得到最佳结果，人类只能在仓促之间找到最佳走法。而计算机的数据库中加载了最佳序列，可以做到这一点。

国际跳棋游戏中，人类击败计算机的唯一机会在中场。人类的预测能力可能不如计算机，计算机会分析不同走法背后大量的可能序列，但人类玩家能更好地把握特定位置的战略价值。例如，人类玩家可能会认识到，控制棋盘中间位置的重要程度比计算机损失函数给出的权重更高，或者计算机控制中间位置的数值测量可能有误，又或者中间位置的最终控制无法依靠测量目前局势得知。

计算机的最后一个优点是它不会累。高水平国际跳棋游戏可以持续两个多小时。由于大多数国际跳棋对决是以平局结束的，因此跳棋锦标赛会安排很多场比赛，一个多星期下来可能每天都会有四场。人类玩家每天的比赛时间为8~10个小时，一天接着一天，他们会疲惫不堪，容易出错。但计算机不会疲倦，因为它不需要思考，只要服从就好。

史上最优秀的国际跳棋选手是传奇人物马里恩·廷斯利。他是一个神童，开始念书的头八年就跳了四级，后来成为专攻组合分析的数学教授。小时候，他每周用五天，这五天每天用八个小时来学习国际跳

棋。读研期间，他称自己已经花了一万个小时研究国际跳棋。到了20多岁，他基本上已无人能敌。

有12年时间，廷斯利不再参加国际跳棋锦标赛，据称是因为他觉得非常保守的对手很无聊——他们希望的最好成绩是平局。后来重返赛场的他于1991年再次退役，时年63岁。1992年他又被国际跳棋程序奇努克（Chinook）团队的创建负责人、数学教授乔纳森·谢弗请回赛场。谢弗的研究团队有三个人，分别负责开盘数据库、残局数据库和中场损失函数。

在1992年廷斯利和奇努克的40场比赛中，大部分是平局。廷斯利赢了第5场比赛，那场比赛中奇努克遵从了已加载在其剧本中的一种次优走法。廷斯利输了第8场比赛，将其归因为疲劳过度。到了第14场比赛，奇努克采用了数据库中廷斯利多年前使用过的一连串走法，但廷斯利忘记了，因此输了比赛。后因奇努克发生故障，廷斯利拿下了第18场比赛。（计算机也会疲劳？）随后，廷斯利还取得了第25场和第39场比赛的胜利，最终以4胜2负33平的成绩取胜。

这是人类大战机器中人类的一次胜利，但两场比赛的失利，却是廷斯利45年国际跳棋职业生涯中仅有的两次。

谢弗极大地扩充了奇努克的开盘和残局数据库，还将中场的前瞻能力从17步增加到了19步。1994年，他要求再进行一次对决。前六场比赛为平局，不过廷斯利认为奇努克的水平已经得到提升。他表示，在奇努克的残局数据库足够巨大以至于不会出错之前，他只有10~12步的机会能获得领先优势。可惜的是，廷斯利因患胰腺癌而不得不放弃比赛，并于7个月后与世长辞。

廷斯利的记忆力惊人。1992年第一次比赛后，他给谢弗讲了自己40多年前的一场比赛，他仍能准确无误地记住每一步。尽管如此，他的记忆力还是无法与强大的计算机匹敌。廷斯利真正拥有的是通过多年研究和实践积累的棋感，奇努克绝不可能对位置的优劣有相同的直觉。

在决胜局前的14场展示赛中，廷斯利和奇努克有13场平局，第10场为廷斯利获胜，谢弗随后描写了这场决定性的比赛：

> 我走下了奇努克的第10步。刚放下棋子，廷斯利就惊讶地抬起头说：“你会后悔的。”我尚未领略过伟大的廷斯利的行事风格，

默默地坐在那里，心想：“你知道什么，我的程序正在搜索后20步的可能性，表示它占优势。”再走几步后，奇努克的评估降至旗鼓相当。又走了几步后，它表示廷斯利更占上风。后来，奇努克说它遇到了麻烦。最后，越下越糟，我们只好投降了。在廷斯利的比赛日志中，他透露自己已经预料到残局，在第11步就知道他会赢，也就是我们出错的下一步。而奇努克需要预测后60步，才能知道它的第10步下错了。

廷斯利去世后，奇努克与世界排名第二的国际跳棋选手唐·拉弗蒂进行了32场比赛，并以1胜31平取胜。1996年，奇努克退出国际跳棋锦标赛，不过你可以在线对战低配版的奇努克。退赛后，奇努克同数十台差不多连续运行了18年的计算机一起工作，以检验确认国际跳棋玩家在先走并且每一步都是最佳走法的情况下是否可以保证取胜。

2007年，谢弗宣布国际跳棋和井字游戏一样，也是一款极好的权衡游戏，如果每个玩家都能选择最佳走法，则可以保证平局。这是计算机的一项壮举，但我不会称其为智能。

下一代的计算机游戏程序采取了不同的做法，即试错过程——计算机跟自己比赛数百万次，同时记录取胜方式。一款名为AlphaGo（阿尔法围棋）的程序采用了这种方法，击败了世界上最顶尖的围棋手。此外，另一款名为AlphaZero（阿尔法零）的程序还击败了最好的计算机国际象棋程序。这些程序都能极好地执行范围狭窄、目标明确的任务（“将”对手的军），但不会像人类那样分析棋盘游戏，思考为什么某些策略会成功。即使是计算机编码员也不明白为什么他们的程序有时会选择不寻常的甚至是奇怪的特定走法。

创建AlphaGo和AlphaZero的公司DeepMind（深度思考）的首席执行官戴密斯·哈萨比斯举了个例子。在一场国际象棋比赛中，AlphaZero将“后”移到棋盘的边角格，这与人类想法相矛盾，因为国际象棋中最厉害的“后”在棋盘中间位置会更加强大。在另一场比赛中，AlphaZero牺牲了“后”和一个“象”，而对人类玩家来说，除非可以立即获得回报，否则几乎不会这样走。哈萨比斯说：“AlphaZero与人类的玩法不同，与编程的玩法也不同。它采用第三种玩法，似乎是外星人般陌生怪异的玩法。”

尽管在棋盘游戏中具有怪异的超人技巧，但计算机程序并不具备类似人类智慧和常识的东西。这些程序不具备处理不熟悉的情况、不明确

的条件、模糊的规则以及含糊甚至相互矛盾的目标所需的一般性智能。决定去哪里吃晚餐、是否接受一份工作、跟谁结婚，都与“象”走三步“将”对方的军截然不同——这就是为什么让计算机程序为我们做决定是危险的，不管它们多擅长棋盘游戏。



第3章
无语境的符号

人类拥有无价的现实世界知识，我们用积累了一辈子的经验来帮助自己认知、理解和预测。而计算机没有这种可以指导自己的现实世界经验，因此，它必须依赖数据库里的统计学模式，这或许会有所帮助，但肯定会出错。

我们使用情绪和逻辑来构建有助于理解所见所闻的概念。看见一只狗，眼前就能出现其他狗的形象，想起猫与狗的相同和不同之处，或料到这只狗会追赶身边的猫。或许我们还记得儿时的宠物，或者回忆起以往遇到狗的经历。想到友好忠诚的狗，我们也许会面露微笑，想摸摸它，或扔根棍子引它追取；想到曾把自己吓得半死的恶狗，我们可能会退避三舍，和它保持距离。

这些都是计算机力所不及的事情。对计算机来说，狗、老虎和XyB3c这种无意义的数字与字母的组合没有太大区别，只不过是不同的符号而已。计算机能统计出一篇故事中“狗”这个词用了几次，检索关于狗的事实情况（如狗有几条腿），但不会像人类那样理解词语，对“狗”这个词也不会出现人类那样的反应。

现实世界经验的缺失，通常在试图解读词语和图像的软件中暴露无遗。

翻译软件与理解语言

语言翻译软件程序可以把某种语言的书面或口头语句，转换成另一种语言的对等语句。20世纪50年代，乔治敦大学和IBM的合作小组展示了机器翻译——利用250个词汇和6项语法规则把60个句子从俄语翻译成英语。该团队的首席科学家预测，输入更大数量的词汇和更多语法规则后，翻译程序在3~5年内就可达到完美。他真是异想天开！他对计算机太过自信了。如今60多年过去了，虽然翻译软件的表现不同凡响，但是仍远远达不到完美。发展路上的绊脚石都具有启发意义。

人类在翻译语句的时候，会先将其放在语境中思考（作者是什么意思），然后用另一种语言表达这一内容。翻译程序并没有考虑语境，因为它们无法理解内容的意思。

翻译程序识别输入语句中的词汇和短语，在已经由人工翻译好的文本数据库中搜索，寻找输出语句的对应词汇和短语。同时还寻求可消除歧义的数据模式。例如，包含baseball（棒球）一词的句子中出现了bat这个名词，该名词含有棒球棒和蝙蝠两种含义。而翻译程序选定最有可能正确的词语后，输出的句子是输出语言按照特定的语法规则构成的。

很多机器翻译程序，包括谷歌翻译，目前都采用深度神经网络（deep neural networks）。这种网络虽然受启发于人脑的神经网络，但并不能模仿人脑，因为我们对人脑是如何运作的探索几乎还停留在表面。深度神经网络比早期的翻译程序更加复杂，听上去也更吸引人，但仍然只是试图匹配词汇和短语，然后连词成句的数学程序而已。和较早的翻译程序一样，当前的深度神经网络每次在翻译语句时，都没有试图去理解作者想表达的意思。

深度神经网络改善了语言翻译（以及视觉识别等很多任务），但还是受限于现实状况。计算机不像人脑，并不能真正理解词汇、图像和生活。无论未来计算机多么强大，即使它能够识别关键词和短语、查找匹配其他语言的词汇和短语、将匹配结果按照语法规则排序，但这些都不算是阅读或写作，与传达意思并非一回事。

机器的翻译速度很快，并且通常都能完成得不错。但有时候也会意思表达不完整，译文令人不解或啼笑皆非。霍夫施塔特给出以下例子：

> In their house, everything comes in pairs. There’s his car and her car, his towels and her towels, and his library and hers.
>
> （在他们的房子里，所有东西都有两份。他有他的车，她也有她的车；他有他的浴巾，她也有她的浴巾；他有他的书房，她也有她的书房。）

霍夫施塔特用谷歌翻译先将这句话翻译成法语，再回译成英语，结果如下：

In their house, everything comes in pairs. There’s his car and his car, his towels and his towels, and his library and his.

（在他们的房子里，所有东西都有两份。他有他的车，他也有他的车；他有他的浴巾，他也有他的浴巾；他有他的书房，他也有他的书房。）

第一句意思明确，译文没问题。第二句却出现了偏差，因为包括法语在内的罗曼语族在语法上有“性”的区分。

不过，问题不仅在于her（她）在译文中没有体现出来。谷歌翻译并不理解（甚至没有想要理解）第二句话是什么意思。通过观察亲戚朋友和自身情况，人们都知道大多数伴侣乐于彼此分享。但这句话告诉我们的是，即便这俩人生活在同一个屋檐下，也宁愿各用各的浴巾、车、书房，（肯定）还有更多其他东西。计算机程序没有生活经历，无法进行这样的观察，也就不知道第二句话想表达的意思，不会试图重现其含义。这并非计算机能力或编程错误的问题，只是反映出一个事实——翻译程序和所有计算机程序一样，无法理解概念和想法。

霍夫施塔特还翻译了卡尔·西格蒙德用德语写下的一段话，请了两名母语为德语的人以及西格蒙德自己来审校译文：

After the defeat, many professors with Pan-Germanistic leanings,who by that time constituted the majority of the faculty, considered it pretty much their duty to protect the institutions of higher learning from“undesirables”. The most likely to be dismissed were young scholars who had not yet earned the right to teach the main university classes. As for female scholars, well, they had no place in the system at all; nothing was clearer than that.

（战争结束后，许多有泛日耳曼倾向的教授认为，保持高等学府不被“不受欢迎的人”侵害是他们的责任。最有可能被开除的是那些尚未获得教授大学主要课程权利的年轻学者。至于女学者，她们在这个体系中根本没有地位；没有什么比这更清楚了。）

将以上由人工翻译的译文与以下谷歌翻译的译文进行比较：

> After the lost war, many German-National professors, meanwhile the majority in the faculty, saw themselves as their duty to keep the universities from the “odd”; Young scientists were most vulnerable before their habilitation. And scientists did not question anyway;There were few of them.
>
> （失败的战争结束后，许多德国国家教授，同时也是教职员工中的大多数，认为自己有责任让大学远离“奇怪”；年轻的科学家在适应训练之前最容易受到伤害。科学家们无论如何也没有质疑；他们很少。）

谷歌翻译的译文几乎让人无法理解，因为谷歌翻译并没有捕捉到文字的意思，它只不过是翻译单独的词汇和短语，然后拼凑在一起。

我推荐大家去看看霍夫施塔特列举的第三个例子，原文为中文语段。谷歌翻译的译文，部分内容曲解了原文的意思，还有部分内容毫无意义。

我之所以反复强调这一点，是因为计算机能够思考的这一想法太诱人了。认为它们能理解世界，提出可靠的建议和决定，这是一种错觉。翻译程序的缺陷充分说明了目前计算机程序的能力与局限。

霍夫施塔特认为：

> 谷歌翻译的开发者无意让谷歌翻译理解语言，而是在想方设法地避开理解需求。他们并不想用文本来模仿构思，只想用语段触发搜索庞大数据库中的其他语段。这就像是“迂回”（end run）战术，以间接方式理解、明白和认识语言的目的。在我看来，这完全自相矛盾、有悖常理。因此，尽管谷歌翻译表面看类似人脑结构，但实际上，其开发者在尽其所能避开人脑可以完成的事情，即理解世界。

这并不意味着计算机永远都不可能模仿人类思维，但如果程序员不做此尝试，或接受“迂回”战术，计算机就不会具备这个能力。我再次引用霍夫施塔特的话，和计算机不同，他能言善道：

> 从原则上说，绝对没有基本性哲学解释证明机器永远不会思考、创造、有趣、怀旧、兴奋、害怕、狂喜、逆来顺受、充满希望。

> 当然，同理可得，没有理由证明机器不能翻译出好的译文。也绝对没有基本性哲学解释证明机器将来无法成功翻译笑话、双关语、漫画书、电影剧本、小说、诗歌，当然还有类似本书的论文。但是，这一切只有在机器能做到像人类一样有生命力、想法、情绪和经历时才能实现。不过，这不会发生在不久的将来。老实说，我认为是遥遥无期的。

威诺格拉德模式挑战赛

斯坦福大学计算机科学教授特里·威诺格拉德参与发起了后来为人所熟知的威诺格拉德模式挑战赛（Winograd Schema Challenge）。以下为纽约大学计算机科学教授欧内斯特·戴维斯编纂收集的一个例子：

> I can’t cut that tree down with that axe; it is too （thick/small）.

> 我没法用这把斧头砍倒那棵树，它太（粗/小）。

如果括号里的词为thick（粗），那么it（它）指的是那棵树；如果括号里的词为small（小），那么it（它）指的就是那把斧头。这类句子——有两个名词，还有可选择的单词表明代词所指的是哪个名词——人类立刻就能理解，但这对计算机来说就非常难了，因为计算机没有现实生活经验来提供理解词汇的语境。

人类根据生活经验会知道如果树太粗或斧头太小，都很难砍倒树。而计算机无法理解这一点，因为它没有生活经验可以借鉴。

著名AI研究者奥伦·埃奇奥尼曾说，计算机就连句子中it的所指都弄不清楚，还怎么谈得上可以主宰世界。

目前，威诺格拉德模式挑战赛设奖金2.5万美元，奖励在威诺格拉德模式下解读准确率达到90%的计算机程序。在2016年的挑战赛中，最高准确率为58%，最低为32%，概率变动更多为运气因素，而非计算程序能力的差异。值得注意的是，谷歌和脸书并未参赛，放弃了一个炫耀自家软件能力的绝佳机会。

计算机能阅读吗？

鲍勃·迪伦荣获诺贝尔文学奖，获奖理由为“在伟大的美国歌曲传统中开创了新的诗歌表达”。他原名为罗伯特·艾伦·齐默尔曼，后随威尔士诗人迪伦·托马斯更名为鲍勃·迪伦。他后来解释说：“你就这样出生了，取了不好的名字，来到了错误的家庭。人生有时就是如此。你可以想怎么称呼自己，就怎么称呼自己。”20世纪60年代，迪伦以抗议歌曲为特色（尤其关于公民权利和越南战争），成为他那个时代的代表声音。

罗杰·尚克作为50多年前开始AI研究的科学家，期望能造出像人类一样思考的计算机，例如，像人类那样理解语句。可事实证明，这个想法极难实现，部分原因是我们并没有真正理解人脑是如何运作的。

20世纪80年代，AI的发展绕道而行，朝商业可行的方向发展，例如，研究词汇（易做），而不是概念（难做）。计算机擅长保存严谨精确的记录和检索信息——这对搜索引擎来说至关重要，但是与认知思维毫无关系。

例如，计算机可以搜索全文查找单词betray（背叛），但无法识别出没有使用betray一词来讲述背叛情节的故事。计算机可以查找单词，但无法理解其意思。2017年，尚克写道：

> 我担心的是IBM关于“沃森”程序的夸张言论。最近，他们发布了一则以鲍勃·迪伦为主角的广告，让我捧腹大笑，或者说，会让我捧腹大笑，如果我没有勃然大怒的话。我想说句大实话：“沃森”就是一场骗局。并不是说它不能处理词汇，对某些人来说，词汇处理能力很有价值。但是，那些广告纯属欺骗。

《广告周刊》的一篇文章指出，“沃森”能每秒阅读8 000万页内容，识别迪伦作品的关键主题，如“时光流逝”和“爱会枯萎”，这证明它和传统编程计算机不一样，像“沃森”一样的认知系统可以理解、推理和学习。

还是让它好好做个单词计数器吧。我不记得迪伦用过civil rights（公民权利）或Vietnam（越南）这些词语（“沃森”肯定不用

一秒就能查到），但是迪伦的歌迷——人类——知道他在20世纪60年代的写作主题是什么——不是“时光流逝”，也不是“爱会枯萎”。

思考一下歌曲《时代在变》（The Times They Are A-Changing）的开头几句歌词：

> 大家集合于此吧
>
> 无论你在何处游走
>
> 承认你四周的潮水
>
> 已经日渐高涨
>
> 承认吧
>
> 不久后你就会被淹没

计算机很容易识别、列举和计算这些词语，但是完全不明白迪伦在说什么。人类或许会对这首抗议歌曲有很多不同的解读（大多数伟大的文学作品都是如此），但是他们的解释肯定远不止停留在识别单个词语上。人类运用词语来表达意思（并不总是直接表达），还利用语境来理解其他人的话语。要计算机掌握这种最基础的人类智能，毫无希望可言。

仔细想想，哪五首是你最喜欢的歌？“沃森”会明白这些歌曲讲的是什么吗？《带我飞向月球》（Fly Me to the Moon）、《自由坠落》（Free Fallin’）、《加州旅馆》（Hotel California）、《生而为逃亡》（Born to Run）、《千载难逢》（Once in a Lifetime）。

计算机能写作吗？

我上高中的儿子在学校打棒球，每场比赛过后，都会在线发布由计算机程序根据比赛记录编写生成的书面总结。以下为克莱蒙高中狼群队对阵钻石吧高中梵天队的比赛总结示例：

> 星期五，狼群一记全垒打，以6：5击败钻石吧。在第八局比赛最后比分为5：5平，狼群的怀亚特·科茨倒地牺牲短打，夹杀得

分。

尽管钻石吧在第二局三次夹杀得分，狼群仍取得了比赛胜利。钻石吧的大局由富勒一垒打、克里斯蒂安·基利安一垒打和费边·莫兰一垒打锁定。

钻石吧在首局开场领先。钻石吧基利安的一记高飞牺牲打击夹杀得分。

狼群在第七局比赛最后将比分扳平至5：5。杰克·金特里击入内野手范围，夹杀得分。

钻石吧在第二局三次夹杀得分。钻石吧的大局由富勒一垒打、基利安一垒打和莫兰一垒打锁定。

［由Narrative Science（自动写作技术公司）和GameChanger Media（移动应用程序和网站）提供支持。版权所有2017年。保留所有权利。］

该总结将钻石吧高中梵天队在第二局中的三次夹杀记录为两次，跳过激烈的赛事直接叙述第八局，又跳到第二局、第一局，再到第七局，最后又回到第二局。称克莱蒙高中狼群队为“狼群”，而不是“克莱蒙高中队”或“狼群队”，这一点也挺尴尬的。虽然这份总结要点突出，但描述枯燥乏味，读者无法从中感受到这场比赛的激动人心之处。从人类的角度来说，更好的总结应该能强调钻石吧高中梵天队开场大比分领先，克莱蒙高中狼群队紧追比分，在第七局末扳平（通常是最后一局）。然后，比赛进入加时决胜局，克莱蒙高中狼群队以自杀式抢分触击反超取胜。我还希望总结里提到，我儿子作为投手，参与了五又三分之一无得分局直到克莱蒙高中狼群队重振雄风！

如今，很多报纸都采用机器撰写文章。《华盛顿邮报》的做法是，编辑将某个主题、主题相关事实发生的地方，以及他们希望在故事中出现的关键词或短语输入计算机程序。该程序拟好一份初稿，编辑在此基础上修改确定终稿。这种做法最适合重事实轻观点的叙事（如棒球赛）新闻和不值得劳驾高薪聘请的作家与编辑下笔的小文章。小镇的报社尤其对此感兴趣，这些报纸的版面内容都是当地新闻，如婚礼、讣告和高中体育活动。

我从中发现了一个很有趣的测试，能比较计算机智能与人类智能——计算机能否写出有说服力的小说。设想在图灵测试中，计算机和小说家都收到提示词，可能是Boston、grandchildren、hardware store、war（波士顿、孙子、五金店、战争）等，被要求使用这些词语写一篇短文。若其他人无法辨别哪篇短文出自计算机，则该计算机通过测试。除了写短文，人类和计算机还可以比赛使用提示词写歌词或诗歌。

多伦多大学的AI研究人员开发出会写歌的“神经系统卡拉OK”（neural karaoke）程序。以下示例受启发于圣诞树的电子照片（我觉得公平地说，那些研究人员之所以宣传这个例子，可能是因为其他计算机生成的歌词情况更糟糕）。

Lots to decorate the room.（有很多东西可以装饰房间。）

The Christmas tree is filled with flowers.（圣诞树上摆满了鲜花。）

I swear it’s Christmas Eve.（我发誓今天是平安夜。）

I hope that’s what you say.（我希望你是这么说的。）

The best Christmas present in the world is a blessing.（世界上最好的圣诞礼物是祝福。）

I’ve always been there for the rest of our lives.（我的余生都在那里度过。）

A hundred and a half hour ago.（100.5个小时以前。）

I’m glad to meet you.（很高兴见到你。）

I can hear the music coming from the hall.（我能听到音乐从大厅里传来。）

A fairy tale.（一个童话。）

A Christmas Tree.（一棵圣诞树。）

There are lots and lots and lots of flowers. （有很多很多很多的花。）

这首歌唱起来比读起来好不到哪儿去。

要想通过创意写作测试，即称得上智能的计算机写作，计算机程序必须知道词语在语境中的意思。这样程序就可以写出令人信服、内容煽情、情节有趣的叙述，能吸引读者继续往下读。目前，计算机还做不到这一点。

InspiroBot

有一个广受欢迎的名为InspiroBot的人工智能海报生成机器人，能巧妙使用一些明显出自人类的词语：

> I’m InspiroBot.
>
> I am an artificial intelligence dedicated to generating unlimited amounts of unique inspirational quotes for endless enrichment of pointless human existence.
>
> （我是InspiroBot。
>
> 我是人工智能，致力于生成无数独具特色的激励语句，为无意义的人类存在增添无限光彩。）

InspiroBot程序有激励信息的常见语句结构数据库，就像聚会时玩的填词游戏，一人选择名词、动词、副词和形容词，另一人将这些词语填入故事的空白处。完成的故事有时搞笑，有时荒谬，因为选词的人并不知道词汇的使用语境。

InspiroBot也是如此。它能把名词放入激励短句中名词的位置，但是它无法知道这句话会激起热情、大笑还是困惑。实际上，计算机生成的信息有可能很空洞，所以该网站得依靠人类假扮机器，写出真正有趣的信息。

以下是InspiroBot为我生成的一些信息：

Where friends radiate, bank robbers melt.（朋友所到之处，银行劫犯消失。）

Embrace greed, remember time.（拥抱贪婪，铭记时间。）

Avoid vegetables and you shall receive a woman. （避开蔬菜，你会得到女人。）

Meditation requires 90 percent love, and 99 percent fake.（冥想需要90%的爱和99%的伪装。）

A believer can be a space alien, but a space alien can also be a believer.（信徒可以是太空外星人，反之亦然。）

If you are the most gentle soul in the laughter, prepare for another laughter.（如果你是笑声中最温和的灵魂，请做好听到其他笑声的准备。）

Breaking the sound barrier makes you go blind, unless you start working out.（打破声音障碍会让你失明，除非你开始锻炼。）

在语境中理解事物

不仅仅是语句中的词语。图像识别程序可将简单图像与计算机数据库中的相似图像进行精准匹配，但若图像出现扭曲、部分模糊不清或内容复杂的情况，就较难为其进行匹配了，因为计算机不能用类比方法识别图片的基本要素。

人类在语境中了解事物。我们在街上开车来到十字路口时，预料可能会看到停车指示牌，自然就会扫视可能会出现指示牌的地方。如果我们见到熟悉的八边形指示牌，上面显示“STOP”（停）的字样，就能一眼识别出来。即使这个指示牌生锈了、凹凸不平或贴着小广告，我们仍能认出它是指示牌。

可是，图像识别软件就无法做到这一点。例如，在研究停车指示牌时，深度神经网络会先扫描不计其数的停车指示牌，识别其共同特

征，再利用这些特征评估某对象是不是停车指示牌。计算机程序不会观察某个对象的通用特征，而会观察独立的像素，通常还会注意到微不足道的特征。AI软件非常靠不住，因为稍有差异就会让软件出错，即便是停车指示牌上有一小张贴纸，也会扰乱计算机的识别。

在训练过程中，深度神经网络会将“停车指示牌”的字样与数不胜数的停车指示牌图像进行匹配，当输入像素与训练记录像素高度相似时，深度神经网络便学会输出“停车指示牌”的字样。无人驾驶汽车在遇到训练标记为“停车指示牌”的匹配像素时，便会自动停车。不过，计算机不明白为什么要停车，也不明白若不停车会有什么后果。人类司机看到被肆意破坏或掉落的停车指示牌也会停车，因为人类能识别出被毁坏的指示牌，也能想到不停车的后果。

关键的问题同上，即AI算法与人脑运作不同。人类不需要看上百万张自行车的图片去了解什么是自行车。就算自行车的把手被系上丝带、车身被粘了闪电的图片，也骗不过人类。

人类识别事物不仅要将其与同类事物进行对比，还要与其他事物进行区分。例如，人脸识别软件研究一张脸，要记录数量惊人的特点，然后尝试将这些特点与计算机数据库中储存的图像的特点进行匹配。该程序不局限于搜索脸部，因为它不知道何为脸部。算法有可能将人脸识别成石头、星球或咖啡杯。

人类的识别方式就不一样，我们想到某个人，也会想到他的脸。人脑的关注点在于这张脸和我们预想的人脸形象——招风耳、瓜子脸、粗眉毛——有何不同，正如讽刺漫画中突出的特色一样。这些差异就是所谓的区别性特征（distinguishing features），人脑能立即识别人脸靠的就是这些差异点，而非相似点。

看到某人缺了颗门牙，我们不是像深度神经网络程序那样注意到他的其他牙齿，而是靠这颗缺失的门牙把此人与他人区分开。同样，帮助我们立刻识别出单车的是我们所看见的两个车轮，而不是3个、4个或18个车轮。帮助我们立刻识别出袋鼠的是，大多数4条腿的动物的前后腿都差不多，不会像袋鼠那样直立，也不会跳着走。

计算机做不到这些，因为它不知道，也不理解这些事物是什么。计算机的方法以颗粒为单位，分析的是像素，而不是概念，所以有时会得到荒唐的结果。

谷歌的一个研究团队表示，人类察觉不到的细微的像素改变都能忽悠最先进的视觉识别程序。他们将这些变化标记为“对抗”（adversarial），说明他们对于捣乱者可能实施的恶作剧心知肚明，例如，对停车指示牌做些难以察觉的手脚来骗过无人驾驶汽车。

怀俄明大学和康奈尔大学的人工智能发展实验室的研究人员展示了更令人惊讶的事情：深度神经网络会把无意义的图片错误解读为实物。例如，将看上去杂乱无章的小圆点和图案识别为海星、猎豹等（如图3.1所示）。

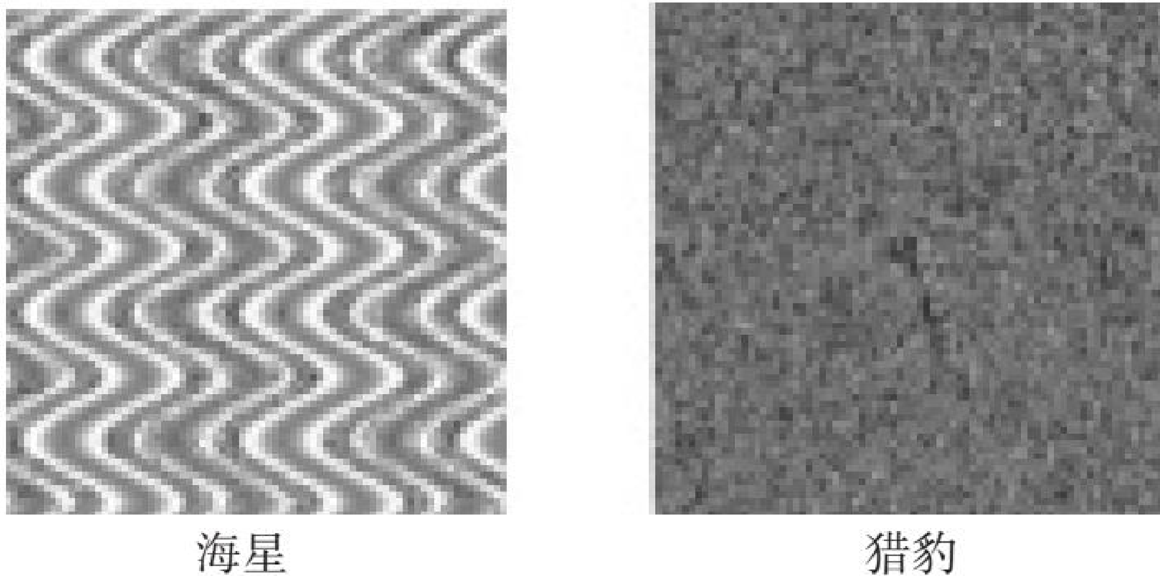


图3.1 无中生有的识别

2016年，另一个计算机科学家团队撰文称，脸部生物识别系统中最先进的深度神经网络程序识别不出戴了有色镜框的人脸。人们不仅能以此隐藏自己的身份，还能通过选择镜框颜色误导系统错误地将其识别为他人。研究者中的一名白人男性被误认为是白人女演员米拉·乔沃维奇，相似度为88%（如图3.2所示）；另一名24岁、来自中东的男性被误认为是43岁的美国电视节目主持人卡森·达利，相似度为100%。这都是因为镜框颜色误导了计算机程序。

人类不会犯这种显而易见的错误，因为我们知道眼镜是什么，并可以不受眼镜干扰看到那个人的脸。



图3.2 哪个是米拉·乔沃维奇？

图像识别和人脸识别系统肯定会有所改善。我只想表明，计算机智能与人类智能相去甚远。人类能够建立联系、理解关系和辨识大局。计算机能处理像素，但不能理解它们所处理的内容。计算机不知道停车指示牌是什么，也不认识猎豹、海星、米拉·乔沃维奇和卡森·达利。

计算机连股票、人和药是什么都不知道，你还会放手让它来选择股票、雇人和开药方吗？

坦克、森林和云朵

美国陆军曾试图采用神经网络识别森林中的伪装坦克。资深研究人员拍摄了200张图片，其中100张为有坦克的森林图片，另100张为无坦克的森林图片，各用其中的一半以“训练”计算机程序区分树木和坦克，其余100张随后被用来验证效果，看看该程序能在多大程度上区分以前没见过的图片中的树木和坦克。结果显示，该程序识别无误。

后来，这个计算机程序被送到五角大楼，但很快就被拒绝了，因为其准确的概率也就和抛硬币差不多。问题在于那些有坦克的图片拍摄于多云天气，无坦克的图片拍摄于晴天。由于计算机不知道自己要找的是什么，因此只关注云朵，而不是坦克。该程序能完美识别出多云天气，但无法识别出坦克。

其实，重点不在于计算机不能辨别出云朵、树和坦克的差异，而在于人类不会犯这样的错误，因为人类知道自己要找的是什么。和人类不一样，计算机无法理解这个世界。

猫与花瓶

走进一间房，看到有只猫坐在桌上，还有一地的花瓶碎片，你立马会猜测可能是猫把桌上的花瓶打翻到地上，摔碎了。你的第一反应也可能有误，或许是人摔坏了花瓶后就离开了，猫不过刚好坐在了原来放花瓶的桌上；或许是一阵风从敞开的窗户吹进来，吹倒了花瓶；又或许是一场地震将花瓶震落在地。

你还可以搜集更多信息来检验自己的推测。还有谁来过这个房间，他会承认是自己打碎了花瓶吗？有多少扇窗是敞开的，外面的风力有多大？最近有地震通报吗？你可能无法得出定论，但是每种猜测都说得通。

计算机也能这样猜测吗？计算机能观察到房间里的一切，甚至可以正确标记大多数东西。但是，尽管它再努力，花再长时间，能像你那样立刻就提出这些猜测吗？它能立刻抛开你绝不会认真考虑的荒谬推测吗？例如，花瓶自己从桌面纵身跃下；椅子飞到桌面上，给了花瓶一巴掌；地毯满屋子飞，撞翻了花瓶。

这是说明人类和机器之间具有根本差异的经典例子。人类会基于逻辑推理和生活观察进行合理猜测。而计算机的综合性思维非常糟糕，例如，它们运用逻辑、模型和证据来理解为什么飞机会飞，为什么夸奖比批评更有用，为什么失业率会出现波动，为什么花瓶会掉下桌面。

人类能将自己从某一领域吸取来的经验教训运用到其他领域。人类记得见过动物打翻东西，从没见过无生命物体自己跳来跳去。人类也很擅长预测日常事件的后果，如在大热天跳入清凉的游泳池、从屋顶跳到水泥车道上、向某人招手、把球踢到窗户上、闭着眼睛骑单车、朝着孩子微笑、对老板大喊大叫，我们很清楚这些情况下会发生什么。

计算机的类比能力极差，也根本无法预计一件事情如何引发另一件事情。计算机没有现实生活认知，这些智慧和常识来自真实生活，储存为记忆中的所读、所见、所思。这就是为什么“大”数据和“大”电脑会制造出“大”麻烦。



第5章
随机性模式

> 每当统计学课程开课的第一天，我都会做超感官知觉（extrasensory perception, ESP）实验。先给学生们展示一枚普通硬币（有时向学生借），然后将其抛投10次。每抛一次，我就刻意把结果印入脑中。与此同时，学生尝试猜测我的想法，然后写下答案。我还会在一张事先设计好的纸上，以圈出H（正面）或T（背面）的方式，记录每次抛投的实际结果，这样一来，学生就无法通过我的手势猜出结果。

谁猜对了10次，谁就能赢得当地一家精品巧克力店的一盒一磅装巧克力。如果你在家也想试试，那就猜猜我在2017年春季的统计课上那10次抛投硬币的结果。我的脑电波或许还留存在某个地方。然后写下你的答案，看看能猜中几次。

抛完10次后，我让学生们举起手来，然后开始公布结果。猜错的学生把手放下，坚持到最后的即可赢得巧克力。曾经出现过一名获胜者，在参与这个游戏的学生人数超过了1 000名后，有人获胜也在预料之中。

我并不相信超感官知觉，所以这个实验的重点并非赢得巧克力。把巧克力设置为奖品，只是为了让学生认真对待这一测试。我的真实意图是想说明大多数人，即便是聪明的大学生，对抛硬币等随机事件也存在误解。这一误解加深了我们的错误想法，即以为电脑发现的数据模式一定都有意义。

早在20世纪30年代，美国的真力时无线电公司（Zenith Radio Corporation）有一档系列节目，每周播出一次超感官知觉实验。无线电广播里的“发送者”随机选择一个圆圈或方框，类似抛硬币，然后想着所选的图形，希望脑海中的图像能传送给数百英里之外的听众。随机进行五轮选择后，听众可以将猜测答案寄给电台。

这些实验虽然不能支持超感官知觉的说法，但确实可以有力证明，人会低估随机数据模式出现的频率。我们大多数人认为，圆圈和方框出

现的次数通常应该相等，而且不会以任何可识别的模式呈现。例如，在一次实验中，121名听众都选择了以下序列：

□ □ ○ □ ○

只有35名听众选择以下序列：

□ ○ □ ○ □

上述两种序列中，都含有3个方框，2个圆圈，但是第一种序列似乎比第二种完美交错的序列更加随机。这么说你同意吗？

只有一名听众选择如下第三种序列，因为大多数人认为随机结果不会这么一致。

□ □ □ □ □

事实上，这三种序列出现的概率完全相等。不过，听众还是不愿意猜有5个方框连续出现，或者两种形状完美交替出现的序列，因为他们觉得这样的情况不会随机发生。你可能也有同样的想法。对了，我在2017年春季课程上的抛硬币结果为：T、T、T、T、T、H、H、T、H、T。你全猜对了吗？

核对完结果，看看是否有人获胜之后，我会让学生数一数自己的答案中连续出现次数最多的是哪个结果。比如以下序列中，连续出现最多的是4次正面：

H T T T [H H H H] T H

而以下序列中，连续出现最多的是3次背面：

H T T H T H [T T T] H

这些序列不像随机结果，但我保证它们就是随机出现的。我抛了20次硬币，得出了以上结果。

过去10年共有263名学生上过这门统计学课，其中报告连续出现4次或4次以上同一面的人只占13%。你的结果是这样吗？

实际上，正面或反面连续出现4次或4次以上的情形，并非完全不可能！在10次抛投中，同一结果连续出现4次或4次以上的概率为47%。我们预计，在这264名学生中，会有124名报告这样的结果，但实际上只有34名。学生都大大低估了同一面连续出现4次、5次，甚至6次的概率。

显然，大家看到正面或反面一直出现都会感到别扭，因为这样的结果不像是随机产生的。连续出现两三次正面后，他们猜反面的念头越来越强烈，以便达到平衡。

不仅统计学课堂上的抛硬币实验如此，在体育比赛、靠运气取胜的游戏和生活中，大多数人也仍未正确认识到随机数据中出现连续情况的概率有多高。因此，一旦出现连续情况，他们的第一反应就是这些数据并非随机得来，其出现肯定有潜在原因，于是就生编硬造出一个所以然。

篮球运动员如果连续投中5次，肯定会“热乎”起来，非常有可能再投中下一球；连续5次选股大赚的金融咨询师必定是金融高手；连续5年势头良好的共同基金一定由金融天才管理。尽管共同基金的表现中唯一具有一致性的是以往业绩无法准确预测将来业绩，但是投资者还是会放弃业绩连年不佳的基金，转投业绩连年良好的基金。

美国国家体育比赛解说员和体育专栏作家名人堂成员梅尔文·德斯拉格，在其最后一篇报刊专栏文章中提及自己在51年的职业生涯中收获的忠告，其中包括一个著名赌徒的建议：“大名鼎鼎的‘希腊人尼克’（Nick the Greek）透露了取胜的秘诀，他训练自己可以持续玩牌八小时而不用上洗手间。按照他的说法，上了牌桌就不应该打断骰气。”唯有低估随机数据连续出现概率的人才会拼命控制膀胱，唯恐“打断骰气”。

我在上文中提出，“在10次抛投中，同一结果连续出现4次或4次以上的概率为47%”。对此，一名活跃的学生表示难以置信，并且编写了计

算机程序来证明我是错的。他编写的程序模拟了100万次抛硬币，并记录每10次中正反面连续出现最多的次数。他的计算机程序也得到了同样的结果。他承认自己的程序证实了我的观点，但他还是不信。他认为，可能是计算机的随机数字生成器出了问题，但他又没那么多时间自己抛100万次来验证。看来，随机数字不会连续出现的想法已经在他的思维中根深蒂固了。

如果抛硬币超过10次，上述概率会更高，连续出现次数会更多。抛1 000次，同一面连续出现大于或等于10次的概率是62%；抛1万次，同一面连续出现大于或等于17次的概率是53%，大于或等于18次的概率是32%。

数据越多，就越能肯定还会产生更多连续出现的结果，以及其他出乎意料的模式。克里斯蒂安·S.卡鲁德和朱塞佩·隆哥合作发表了一篇理论性文章，题为“大数据中假性相关的泛滥”（The Deluge of Spurious Correlations in Big Data），表明在所有庞大数据中集中出现高度规则的模式都不足为奇。不仅如此，而且：

> 数据越多，就越会在其中发现随意、无意义和（对未来行动）无作用的相关系数。因此，自相矛盾的是，我们得到的信息越多，就越难从中提取有意义的发现。信息量过犹不及。
>
> 如果存在一组有助于做出预测的真实统计学关系的固定数据集，数据滥用肯定会提高无用统计学关系在真实关系中的比率。

假设股价、失业率和利率之间存在因果关系。如果失业率上升，则股价下跌。如果利率上升，则股价也呈下滑趋势。通过看股价、失业率和利率的数据，我们可能会找到证实这些因果关系的统计学证据。

再假设，我们把几座偏僻城市的日常气温也考虑在内，尽管它们跟股价毫不相关。根据卡鲁德和隆哥的论证，纳入的无关变量越多，就越能肯定得到的是无意义模式。

与包含两个有意义变量（失业率和利率）和100个无意义变量（100个小镇的气温）相比，包含两个有意义变量和5个无意义变量的结果可能与股价的相关性更高。与包含两个有意义变量和1 000个无意义变量相比，包含两个有意义变量和50个无意义变量的结果可能与股价的相关性更高。

因此，卡鲁德和隆哥总结道：“数据越多，发现无意义模式的概率就越高。”

数据挖掘

人工智能是不断变化的专有名词，包括计算机模拟人类行为的各种活动，例如，组装汽车、识别物体、将语音转换成文本。人工智能还可以驾车、下棋和交易股票。

控制人工智能活动的计算机程序被称作“算法”（algorithms），即完成任务所需的分步规则。例如，寻找某数平方根的算法步骤如表5.1所示。

算法在进行了5个循环后，得出答案为X=7.071068。

表5.1 平方根算法

规则	步骤
1. 输入任意数 Y	Y = 50
2. 选择测试方程式 X = Y/2	X = 50/2 = 25
3. 计算 X 的平方	$X^2 = 25 \times 25 = 625$
4. 计算 $Z = Y - X^2$	Z = 50 − 625 = −575
5. 计算 E = Z/Y	E = −575/50 = −11.5
6. 若 $\|E\| < 0.00001$，得到 X；否则，进行第 7 步	进行第 7 步
7. Z/(2X) 加上 X	X = 25−575/50 = 13.5
8. 返回第 3 步	进行第 3 步

计算机程序使用多种语言执行算法。平方根算法可以用BASIC、Java、C++等计算机编程语言。当然，人工智能算法的能力远不止这个简单的

例子。

数据挖掘可能是最艰巨、最危险的人工智能形式。传统的数据统计学分析遵从已经广为人知的科学方法，用科学知识取代迷信。研究人员基于观察或推测提出问题，比如，“维生素C是否会降低普通感冒的发病率和严重程度”，研究人员搜集数据后，最好能够通过控制实验来验证这个推测。如果服用安慰剂和维生素C的结果出现令人信服的统计学差异，则这项研究得出结论，维生素C具有统计学上的显著影响。该研究人员运用数据验证了推测。

数据挖掘则另辟蹊径，其数据分析不会受到预先形成的推测的驱使或妨碍。数据挖掘算法的编程目的是发现数据的走势、相关系数等模型。一旦发现有意思的模型，研究人员就创造理论来解释它。或者，研究人员认为，数据可以自圆其说，一切解释都包含在数据中。他们不需要理论学说，只要有数据就足够了。

在维生素C的例子中，假设数据挖掘工具针对1 000个人创建数据库，记录他们的所有信息，如性别、年龄、种族、收入、发色、瞳孔颜色、就医记录、运动和饮食习惯等。接着，使用数据挖掘软件识别出与个人患病天数在统计学上最相关的五项个人特征。结果可能显示为：酸奶食用过量、茶类饮用不足、喜欢散步、绿瞳孔，以及在脸书上最常使用的词为excellent（好极了）。

数据挖掘工具可能得出结论——酸奶、茶、散步、绿瞳孔和脸书常用词为excellent代表着不健康——于是编造出稀奇的故事来解释这些相关系数。数据挖掘工具还可能认为，数据已经解释得面面俱到，无须进一步解释了。

《经济学人》在2015年发表的题为“与悲观相去甚远：经济学发展”（A Long Way From Dismal: Economics Evolves）的文章指出，（研究失业、通胀等的）宏观经济学家应该效仿在科技企业从事产品、公司和市场相关数据挖掘工作的微观经济学家。

> （宏观经济学家）应该减少理论空谈。宏观经济学家都是严谨之人，先创建理论模型，后使用数据检验。新一代经济学家则忽略白板功能，只集中处理数据，让计算机识别出模式。

《经济学人》是一本优秀的杂志，但不是优秀的新闻报道。

2008年，美国《连线》杂志总编辑克里斯·安德森撰写了一篇引起争议的文章，题为“理论的终结：数据泛滥使科学方法过时”（The End of Theory: The Data Deluge Makes the Scientific Method Obsolete）。安德森表示：

> 只要有足够多的数据，数据就能自圆其说……更庞大的数据以及处理数据的统计学工具，都为理解世界提供了全新的方式。相关系数可以取代因果关系，科学的发展根本无须相关模型、统一理论或任何真正的机械论的解释。

当时看来，这似乎是一种刻意煽动争议、几乎毫不掩饰的自吹自擂——“未来是大数据和大电脑的世界，请阅读《连线》”。

值得赞扬的是，数年后，《连线》杂志的英国版发表了一篇具有警戒意义的文章，题为“如何篡改统计值”（How to massage statistics），其中谈到了我的担忧——“计算机让摆弄数据更加轻而易举”，还列举了篡改、挑拣和破坏数据以造成误导的各种方法。

不幸的是，对曾经颇有争议的事情，人们现在已经习以为常。认为处理数据便足矣的人比比皆是——认为人类无须理解世界，也无须理论，能在数据中找到模式就足够了。在这个方面，计算机可谓得心应手。因此，我们应该将决定权交给计算机。

有时，“数据挖掘”这个词的使用范围更广，还包括搜索引擎和机器人汽车工等大有裨益、无可厚非的活动。我经常使用“数据挖掘”来描述这种做法——运用数据发现统计学关系，然后以此预测行为，例如，寻找统计学模型以预测汽车采购、贷款拖欠、患病或股价变动的情况。

知识发现

我和一名教“知识发现”这门课程的教授吃过午餐。我问他，假如缺乏理论（或常识），我们怎么知道由数据产生的模型真的有助于预测，而不是偶然？他认为：

> 证据就在数据之中。我们不仅不需要理论，理论化还会限制我们所见，妨碍我们发现意料之外的模型和关系。模型是否有用，只

> 需要看数据就知道了。这就是为什么我把这门数据挖掘课程称作“知识发现”。

数据挖掘还被称为“数据探索”“数据驱动的发现”“知识提取”“信息获取”等，这些称呼都反映了一个核心思想——数据先于理论，甚至通常无须理论。

很多被称作人工智能的事物都令人惊叹。可是，数据挖掘并非如此。其根本原因很简单，却不易被认识到：

> 我们以为模型不同寻常，因此具有意义。

> 在大数据中，模型无法避免，因此毫无意义。

黑匣子

我最近看了一家对冲基金（我称其为“想都不想”）的企划书，其中吹嘘道：

> 我们完全自动化的投资组合按照计算机算法运行。所有交易均通过复杂的计算机系统完成，消除了经理人的一切主观因素。

这就是所谓的“黑匣子”方法，把内容输入算法，算法输出结果（如图5.1所示），而人类用户对结果的决策过程一无所知。



图5.1 黑匣子

在求平方根的算法中，如果输入50，则输出7.071068。然而，我的算法不是黑匣子，因为我解释了程序是如何运行的，任何人都能检查我在逻辑或某一步指令上是否犯了错。事实上，你可能已经发现了问题。50的平方根可以是+7.071068或-7.071068，而我的算法结果只显示了正数。此外，该程序在求Y=0的平方根时会出现问题，因为第五步是计算Z/Y，但是Z/0无意义。最后，算法如何处理负数的平方根呢？没法处理。

当程序处于开放状态时，人类能够看到运行过程，查找错误、遗漏和其他故障。但当程序藏在黑匣子里时，人类就无法这么做了。我们不知道黑匣子里的算法是什么，无法评估过程中是否存在逻辑错误、编程差错或其他问题。黑匣子的输入内容不计其数，处理过程神秘莫测，输出内容让人难以捉摸。

对黑匣子股票交易算法来说，输入值可能是股价、交易股票数量、利率、失业率、推特出现“股市”一词的次数、黄色涂料的销量和几十项其他变量，输出值可能是100股苹果公司股票的买卖决定。

使用黑匣子交易算法决定股票交易的用户不知道做这些决定的理由，也并不费心去了解，因为他们相信黑匣子，就像希拉里·克林顿相信“阿达”一样。他们认为，计算机比自己聪明，这应该让人放心。包括“想都不想”的对冲基金经理在内的许多人都认为，用黑匣子进行投资决定，这是特点，不是缺点，毕竟它“消除了经理人的一切主观因素”。

核准贷款的黑匣子算法拒绝贷款申请的理由可能是申请人的手机没充满电，监狱假释的黑匣子算法拒绝假释的理由可能是申请人戴着宽腕套，防止犯罪的黑匣子算法建议抓捕某人的理由可能是他的鼻子和嘴巴呈某种形状。你可能觉得我是在胡编乱造，可我说的都是实话。

计算机算法能连续无误地进行数学计算，是因为软件工程师确切地知道自己想要算法去做什么，然后编程实现这一目的。但数据挖掘算法就无法这么做，因为该算法的意图模糊不定，结果无法预测。一名人工智能专家写道：“任何两种人工智能设计之间的相似点，可能比你和矮牵牛花之间的还少。”

黑匣子数据挖掘是人工操作，但它并不智能。这就是为什么我给本书取名为《错觉：AI如何通过数据挖掘误导我们》。

有些人使用贬义词“人工蠢能”（artificial stupidity，AS）来描述计算机让我们失望时的情况，如Siri听不懂问题、谷歌地图导航进了死胡同、自动交通灯卡在红灯上。我使用“人工低能”（artificial unintelligence），并非描述计算机偶尔会犯错误，而是强调计算机并不拥有人类般的智能。为了计算50的平方根而遵守规则，同知道苹果公司股价和墨尔本高温的意义并明白为什么两者之间不存在逻辑关联有根本区别。

大数据、大电脑、大麻烦

几十年前，数据匮乏，计算机还没出现，研究人员奋力搜集数据，并花费数小时甚至数天时间埋头苦算。如今，我们生活在大数据的时代，计算机可以高速运行，二者的有力结合一直受到称赞，甚至崇拜。有些人服从计算机，认为计算机无所不能。对大数据的崇拜被称为“数据主义”（dataism）或“数据化”（dataification），认为一切重要事物都可以用数据来表示，数据分析无懈可击。向计算机臣服吧！

这种痴迷并非没有危害。我们过于武断地认为搜索处理堆积如山的数据不会出差错，但出错在所难免。数据不过是数据，计算机也不过是计算机。计算机无法区分有用数据和无用数据，无法分辨合理结论和一派胡言。“数据无须理论支撑”是一种危险的理念。

连续出现、相关系数、走势模型等本身证明不了什么。即便是通过抛硬币，也能发现这些模型。我们需要思考原因，要问为什么，而非是什么。

不可否认，计算机令人惊叹，神秘莫测。我们大多数人不了解手机如何让我们与几千英里外的人视频对话，也不知道计算机如何能给出详细的驾驶导航，还可以根据当前交通状况给出预计到达时间。我们只知道计算机太神奇了。如果计算机告诉我们，总统大选的结果可以通过闻所未闻的几座城市的气温预测得到，我们可能也会认为它说得对。如果计算机可以显示π的小数点后2 000位数和世界上每座城市的街景图，我们区区凡人有谁能质疑它的智慧呢？

事实的残酷在于，数据挖掘算法是由数学家创建的，相比现实状况，他们对数学理论更感兴趣。从15名数学家的脑部功能性磁共振成像图

（fMRI）可以发现，看到数学等式会激活他们的眶额部皮质中线部，而在人们看到惊险杂技或听到美妙音乐时，这一区域也会激活。有些人欣赏优美的音乐、艺术、舞蹈和文学，而数学家则欣赏数学等式的内在美。

沃伦·巴菲特曾发出警告："要小心满脑子都是公式的怪人。"我大学主修数学，现在教金融学和统计学。实际上，我生活中的每一天都会用到数学，还编过几十个软件程序，为我的研究分析数据。我很喜欢公式和计算机，但我也知道，数学的魅力会引导我们创建让内心愉悦却无实践价值的数学模型。有太多数据挖掘算法都属于这一类。

利益冲突

哪里有利可图，哪里就有人蜂拥而至。

20世纪90年代，计算机进入我们的生活，互联网的发展催生出数以百计以互联网为基础的企业，即广为人知的网络公司（dot-coms）。有些网络公司有好的想法，逐渐发展成为实力雄厚的成功企业，但大多数没有。有太多网络公司只是为了在公司名称里加上dot-com，然后转卖出去，赚得盆满钵满后转身就走。找到好点子、开公司、打造成功企业，然后托付给子孙后代，这是旧经济的过时做法。

一项研究发现，企业不过是在名称里加上了.com、.net或互联网，股价便翻了一番还多。股民的钱打了水漂！

如今，人工智能同样如此。人工智能已经成为一种时尚，任何跟计算机沾边的东西似乎都能被称作人工智能。真可笑，连我那个求平方根的计算器都能被算作人工智能。何乐而不为？

我从前的一名学生投资了人工智能初创公司，他跟我说："当前，'数据科学家'和'机器学习专家'是最热门的职业。其中有些是接受过训练的统计学家、经济学家，但有些只是上过六周网络课程的程序员，课程可能仅重点讲解一些技术工具和技巧，没有提供基础的理论知识帮助他们了解理论的局限。"谁还愿意思考呢？都冠以人工智能的名号，四处兜售就好了。2017年，"AI"入选美国全国广告商协会的"年度营销词"。

我的另一名学生现在是一家大公司的首席财务官，他写信给我：“你不会相信人们多么频繁地向我提到‘大数据’的优势，或者愿意提供‘分析专长’——这些人都是外行，（有可能）没有意识到你在书中详述过的局限。”

为了说服大家为实际上并不需要的东西砸更多钱，需要做出更多承诺，提供超出实际能兑现的范围的目标。这种情况在互联网泡沫时期出现过，如今到了人工智能时代又重蹈覆辙。我们应该对拼命向我们推销的人持怀疑态度。

天生就会被骗

人类不太能接受“随机事件”，见不得某件事无缘无故地发生。我们老想着给每个模型做出有意义的解释，但有可能它根本就毫无意义可言，不过是偶然发生的罢了。正如尤吉·贝拉所言：“这巧合得太不像话了。”

你可以将此怪罪于我们的远古祖先曾设法应对的演化和环境问题。拥有便于生存繁殖的遗传特征的有机体，会将这些特征遗传给后代，而那些欠佳的特征则会被淘汰出基因库。持续不断地代代相传，这些有价值的遗传特征便会占据主导地位。

识别和解释模式曾经具有生存价值。乌云通常预示着下雨，灌木丛中传来声音说明可能有捕食者，发质是繁殖力的象征，脸型对称代表基因健康。远古时期，模式识别有助于人类祖先找到食物和水、意识到危险，还有助于吸引到有繁殖力、能养育健康后代的配偶，并将这种能力遗传给后代。那些不太擅长识别有益于生存繁殖的模型的人，将自己的基因遗传下去的机会更少。经过无数代自然选择，我们天生就会寻找模型，并为找到的模型寻求解释。

我们太容易被内在欲望所诱惑，想要解释所见的事物，这掩盖了如下事实：模型不可避免地是由无法解释的随机事件创建出来的，如抛10次硬币。我们应该承认自己容易受到模型的诱惑，从而努力做到拒绝诱惑，保持质疑。

为模型所惑

真力时公司的超感官知觉测试说明，我们对随机数据有先入为主的想法（或误解）。随机数据看似序列1：

□ □ ○ □ ○

随机数据不像序列2：

□ ○ □ ○ □

随机数据肯定也不像序列3：

□ □ □ □ □

因此，我们认为，如果模型如序列2和序列3，肯定就不是随机产生的。或许是方框和圆圈没有被预先打乱，又或许这不是超感官知觉测试，而是公开播放给间谍的密码。

听到这儿，你或许只是付之一笑，但担任《纽约时报》金融专栏作家多年的伯顿·克兰曾说：

> 我一直都深信不疑的是，（曾经用来记录股价的）纸带上价格之间的点是密码（如图5.2所示），以便互相发出市场波动的信号。有人甚至让我看过所谓的翻译码。

IBM......T........ADM.....X.......ASR......GE......GM......
....222 $\frac{3}{4}$..232 $\frac{5}{8}$....30 $\frac{1}{2}$...192 $\frac{1}{4}$.....7 $\frac{3}{8}$.....331....75 $\frac{1}{2}$.

图5.2 用来记录股价的纸带

破译纸带上随机出现的点是数据挖掘的雏形：先寻找模型，然后为此编一种说法。偏执的股票交易员确实会仔细查看这些点，找寻模型，发现模型，然后设法将这些模型和股价变动联系起来。交易员受模型驱使，拼命寻找模型，而且成功了。他们并没有意识到，模型肯定会出现，即便在随机产生的数据中也一样。

这一误解的另一表现是一本关于如何赢得掷双骰子游戏的书。作者在拉斯维加斯一家赌场记录了5万次掷骰子的结果，研究数字出现的序

列。预计掷骰子5万次会出现约20次4-4-11序列，但实际该序列出现了31次。于是该书建议每当4连续出现两次后，都要押11。作者还发现，38次掷骰子的结果中，7-12-7的序列出现了10次，接着出现的数字为2、3或12。如果这38次每次都押注100美元，就能赢4 200美元。

这些计算都是手工完成的，那时还没有计算机，更别说数据挖掘软件了。一想到作者要花上好几个月甚至好几年来找寻这些模型，我就不寒而栗。唯一令人感到欣慰的是，作者在研究数字上花的时间越多，在偶然事件上押注的时间就越少。

这个可怜的作者为了在5万次掷骰子中寻找偶然模型而耗费时间，今天的数据挖掘软件也在这么做，只不过计算机化的数据挖掘能在数秒内完成这项任务，无须数月时间。掷骰子是简单易懂的例子，它说明了如何总能在这样随机产生的数据中找出模型，以及人们多么渴望自己找到的模型是有意义的。事实上，找到的模型根本毫无意义。

随机噪声

大电脑搜遍大数据后，一定可以找出比掷骰子的4-4-11序列更复杂、更不寻常的模型，即便这些数据只是随机噪声。例如，我为100个随机产生的变量均创建250个观察结果，每个变量初始值为50，在随后的249次观察中，由计算机的随机数字生成器决定这个值是增加还是减少。这100个变量都通过统计学家称为“随机游走”（random walk）的程序产生，就像醉汉走路，每走一步都和前一步没有关联一样，每个变量的下一次改变都与上一次改变没有关系。

每个变量的每次观察都与其他99个变量的演变完全独立开来。但事实上，还是一定会出现偶然性模型。数据挖掘软件有非常强大的模型寻找能力，不过，对模型评估就无计可施了。就像我们在前面章节反复说过的，原因在于计算机并不能理解真实世界。数字只是数字。

我运用某些数据挖掘软件发现这些随机产生的变量中，有一个变量连续13次出现增加情况。如果不是头脑清醒，我可能会认为自己有什么重大发现了。

接下来我用数据挖掘软件寻找任意两个变量之间简单的两两相关系数。一共存在4 950对可能的相关系数。我的数据挖掘软件找到了98对

相关系数在0.9以上的变量。如果不是头脑清醒，我可能会认为自己又有什么重大发现了。

最后，我使用数据挖掘软件来寻找这100个解释性变量中的组合，该组合会与一个真实变量高度相关，即2015年标准普尔500指数的每日价值。每5个变量一组，则有75 287 520种可能。这听上去似乎很多，但是对现代计算机来说不算什么。据我预计，在这些虚假变量中，某些变量组合会与真实变量高度相关。结果不出我所料，数据挖掘软件找到一个组合，与标准普尔500指数的相关系数达到0.88。如果不是头脑清醒，我可能会认为自己真的有什么重大发现了。

数据挖掘软件每次都能发现模型，某一次它表明，精明老练的投资人会战胜股市。该软件会筛选、分类和分析所有随机数据，尽管这些数据跟股价一点关系都没有，对决定买进还是卖出股票完全没有帮助。不过，该软件还是找到了足够强的相关系数，说服黑匣子股票交易算法买进或卖出股票。

只要了解数据是如何产生的，人类立刻就能理解这个笑话，但计算机不能。数据挖掘软件无从明白自己的发现是否有用，因为对计算机来说，数字只是数字而已。

真正进行数据挖掘的人会在大数据中启动其数据挖掘算法，通常是数十亿或数万亿次，他们的算法不仅在每个数据组合中寻找模型、寻找不同数据组合之间的交互关系，还会寻找更加复杂的关系。他们必然会找到不同寻常的模型，不过，就像上述的股市例子一样，软件无法辨别何为因果、何为偶然。

业余的天气预测

再举一个例子说明数据挖掘的危害之处。即使并没有充分的理由表明数据具备实际价值，但数据挖掘工具也照例筛选与预测对象毫无关联的数据。例如，假设我想预测明天的气温。真正的天气预报会使用复杂的计算机模型，将大气分为若干个立方体，运用卫星数据估算每个立方体的气温、湿度、风速等。计算机模型利用物理学、流体力学等科学原理，预测天气会如何随着立方体之间的相互作用进行变化。

这听上去挺费劲的。我没有那些资源，也不懂科学原理。但是，我可以使用数据挖掘软件基于知识发现来预测天气。具体来说，我尝试根

据城市B昨天的气温，来预测城市A明天的气温。我也可以参考城市A昨天的气温，但这就不算是知识发现了，不是吗？

我请一名出色的研究助理海蒂·阿蒂格帮忙搜集了25座分布广泛且相对偏僻的美国城市在2015年和2016年的每日最高和最低气温数据。她无意中把澳大利亚西部一座临时小型机场——科廷机场也包括在内。

真是无巧不成书。几年前的圣诞假期，我到澳大利亚墨尔本拜访朋友。那时，我了解到了葡萄干布丁、澳大利亚的圣诞歌曲和墨尔本板球场的节礼日板球赛，我在后院还把网球当板球打。然而，我印象最深的还是拆圣诞礼物的时候。两兄弟给年迈的母亲送了去西澳首府珀斯的往返机票。母亲打开信封，眯着眼睛看着机票，皱着眉大声抱怨道："我为什么要大老远跑去珀斯？"她住在墨尔本，位于东澳，一辈子都没有飞越过整片国土跑到西澳度假，也没兴趣这么做。

为了纪念这次旅行，我把科廷作为预测城市，看看通过24座同样偏远的美国城市的每日最高和最低气温，运用数据挖掘软件来预测科廷每日最低气温的准确率有多高。我的数据挖掘查到华盛顿州的奥玛克，这是一座冬冷夏热的美国小城市，常住居民不足5 000人。其当日最高气温与西澳科廷机场次日最低气温的相关系数为-0.77（如图5.3所示）。

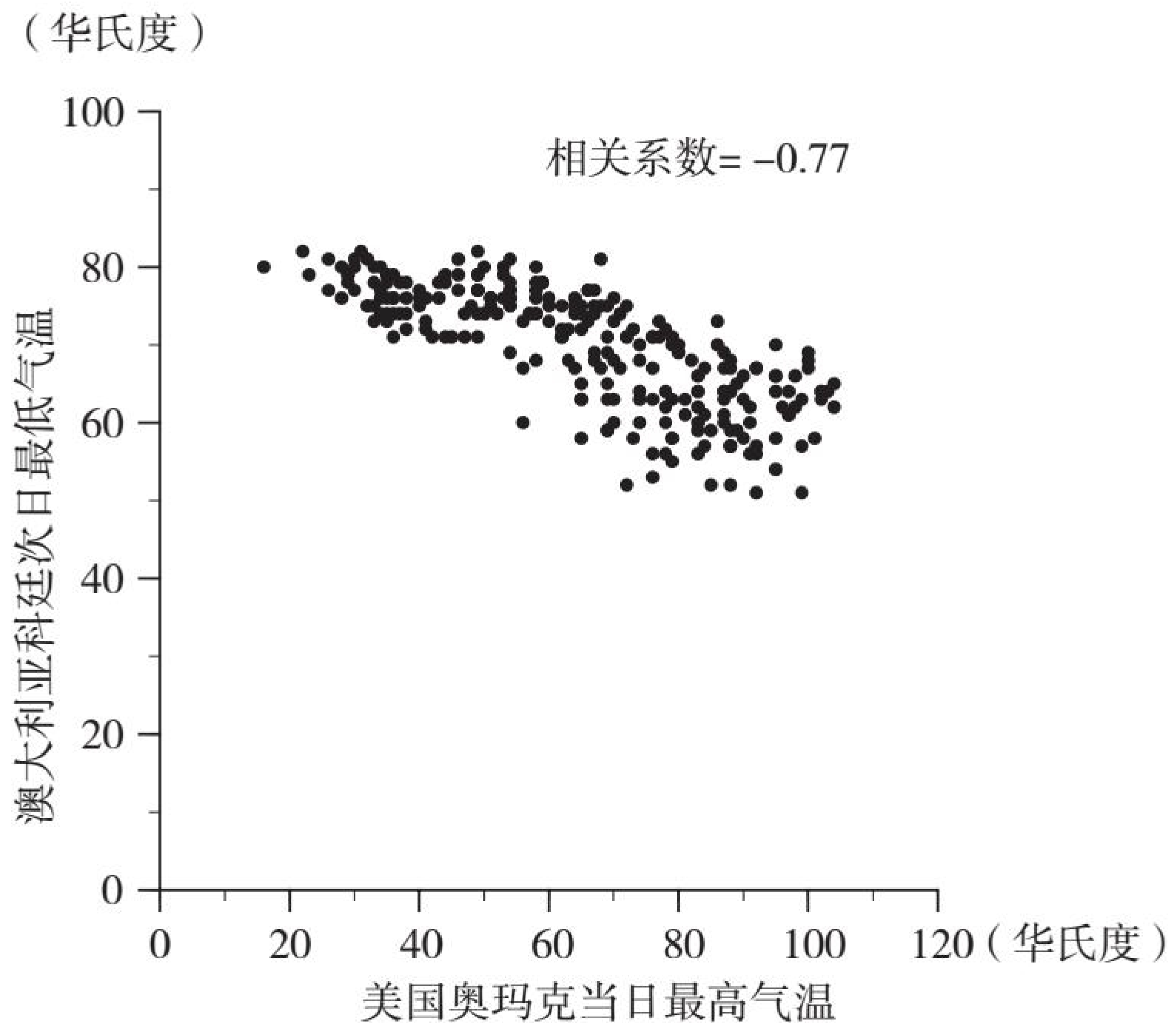


图5.3 根据奥玛克来预测科廷天气的散点图

奥玛克的当日最高气温与科廷次日最低气温呈负相关关系，是因为奥玛克位于北半球，科廷位于南半球。考虑到这两个城市位于不同半球，-0.77的相关系数非常令人震惊。

不会思考的数据挖掘程序（所有数据挖掘程序都不会思考）可能会得出结论，这是一次知识发现，为预测科廷的气温找到了有力的工具。而在会思考的人类看来，预测澳大利亚一个小镇次日最低气温的最佳方法居然是根据远在华盛顿的一个小镇的当日最高气温，这简直荒唐可笑。

我在搜集的另一组数据中启动数据挖掘软件，很快发现了更加紧密的相关系数。如图5.4所示，科廷的每日最低气温与第58号随机变量的相关系数为0.81。没错，图中横轴的变量就是我用计算机随机数字生成器得到的、预测股价的那100个变量之一。

这些虚假变量的生成完全与科廷的天气无关，但我还是发现了一个变量（第58号随机变量）恰好与科廷的天气紧密相关。这就像抛硬币和其他随机噪声那样，通常都会得到看似真实但实则毫无意义的模型和相关系数。

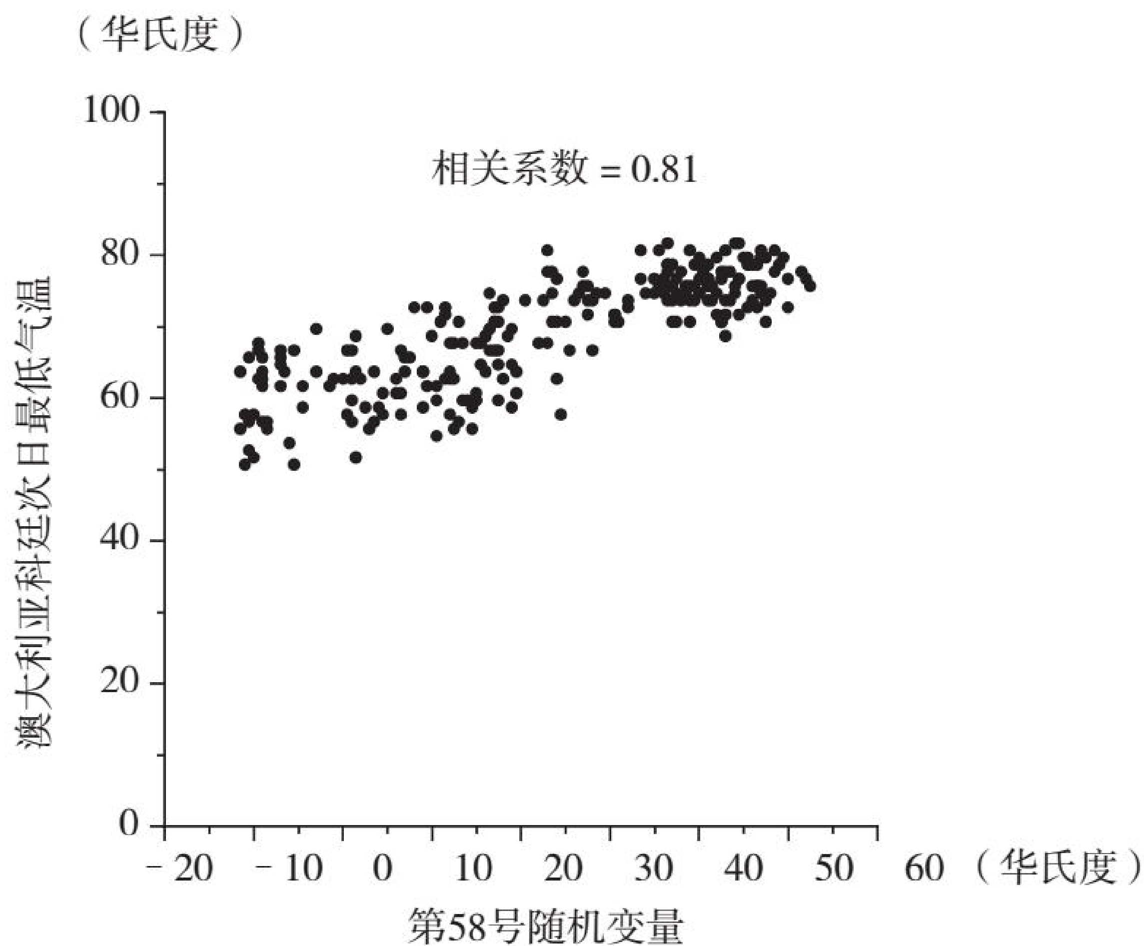


图5.4 随机选择的力量

我这才试了100个随机变量。有了现代计算机，我还可以轻而易举地尝试数千、数百万个随机变量，直至偶然发现一个与科廷或其他城市的

气温存在极其紧密相关关系的系数。

那么，我到底证明了什么？根本什么都证明不了。这就是关于数据挖掘需要记住的第一点，无论是否存在真实情况，只要仔细审查大量数据，就能得到统计学模型。此外，即便被称作人工智能，数据挖掘软件也不足以智能到分辨反映出真实关系和偶然关系的模型有何不同，这唯独人类能做到。

史密斯测试

假设数据挖掘算法发现美国股价与澳大利亚科廷的每日最低气温相关。计算机程序怎么会知道这一统计学关系是真实存在的还是偶然的呢？相反，人类知道何为股价，何为气温，还知道股价高低不由科廷的气温来决定。

计算机能搜索stock的定义，尽管该词有多项词义，如存货、家畜和肉汤等。计算机即便能找出正确的定义，也不知道这个定义中用到的词语是什么含义，虽然它还能继续搜索到定义中每个词语的定义。除了搜索定义外，计算机无法知道股票、股票交易和股价真正代表什么，也不知道为什么股价会时涨时跌。它不明白科廷的最低气温为何物，也不明白为什么这些气温有可能或不可能与美国股价相关。

计算机程序可以搜遍已发表的研究数据库，寻找提及股价与澳大利亚气温的文章。但是对计算机来说，要解释碰巧包含这些词语的研究的相关性，则是难于登天（或是无稽之谈）。此外，评定研究是否有效，对计算机来说也是难上加难。约翰·约安尼季斯曾很有说服力地指出：“大多数已发表的医学研究都有误，包括发表在最负名望的医学杂志上的研究（因为报告结果通常都通过数据挖掘的方法获得）。”我相信，大多数的股市研究也一样。我们会在后面的章节中探讨这些论点背后的推理过程；目前的重点在于，用计算机搜索词语“股价”和“澳大利亚气温”，不可能找出任何被它自己解释为支持或反对其发现的统计学模式的内容。就算确实有所发现，计算机也很难评估其可靠性。

另外，“知识发现”的整套言论都在说，计算机会发现崭新的、从不为人所知的模型和关系。根据这一定义，“知识发现”并非已经发表的事物。那么，没有智慧和常识的计算机又如何能辨别出它的“知识

发现”是否合理呢？它做不到，因为计算机确实没有智慧，也没有常识。

我们回到前述的汉语室测试。如果计算机不能真正理解“股价”和“气温”在现实生活中所代表的意思，那么它就不能分辨出其发现的统计学模型是有意义的，抑或只是巧合而已。可以将这种分辨能力称为“理论性知识”“人类本能”“经验”“智慧”“常识”，不过，通过数据发现统计学关系的计算机和无须数据就能预测关系的人类之间，存在根本差别。

我斗胆提出史密斯测试：

> 搜集100套数据，例如，美国股价、失业率、利率和米价、新西兰蓝色涂料的售价，以及澳大利亚科廷的气温等数据。让计算机自由分析，然后报告它认为可能有助于预测的统计学关系。如果人类专家小组一致认为计算机选择的关系合理，则计算机通过史密斯测试。

有可能存在真正的“知识发现”，即计算机能找到人类忽略的合理关系。但是，如果计算机选择的关系被人类认定为无意义，如美国股价和澳大利亚科廷的气温之间的关系，则其无法通过测试。



第7章
无所不包的“厨房水槽法”

> 20世纪80年代，我曾与一名经济学教授交谈，他根据图7.1所示的简单相关系数给一家大银行进行预测。如果想要预测消费性支出，他便制作收入和支出的散点图，然后用透明尺子画出一条似乎与数据一致的线。根据他的预测，若收入增加，支出也会增加。

这名教授的散点图的问题在于，世界并非如此简单。收入会影响支出，财富状况也会。如果教授恰巧利用收入增加（支出增加）、股市暴跌（支出减少）时期的数据来画散点图，而财富的影响力又大于收入的影响力从而导致支出减少（如图7.2所示）那又会怎么样呢？据此，教授的收入和支出散点图将预测：收入增加，则支出减少。之后，当他试图预测在某一收入和财富都增加的时期支出的变化趋势时，他会预测到支出呈下降趋势，这简直错得离谱。

此时需要运用多元回归分析。

多元回归模型含有多个解释变量。例如，消费性支出模型可表示为：

$$C= a+bY+cW$$

C代表消费性支出，Y代表家庭收入，W代表财富状况。

以上解释变量的罗列顺序并不重要。重要的是将哪些变量纳入该模型，哪些排除在外。回归分析的技巧重点在于选择重要的解释变量，忽略不重要的解释变量。

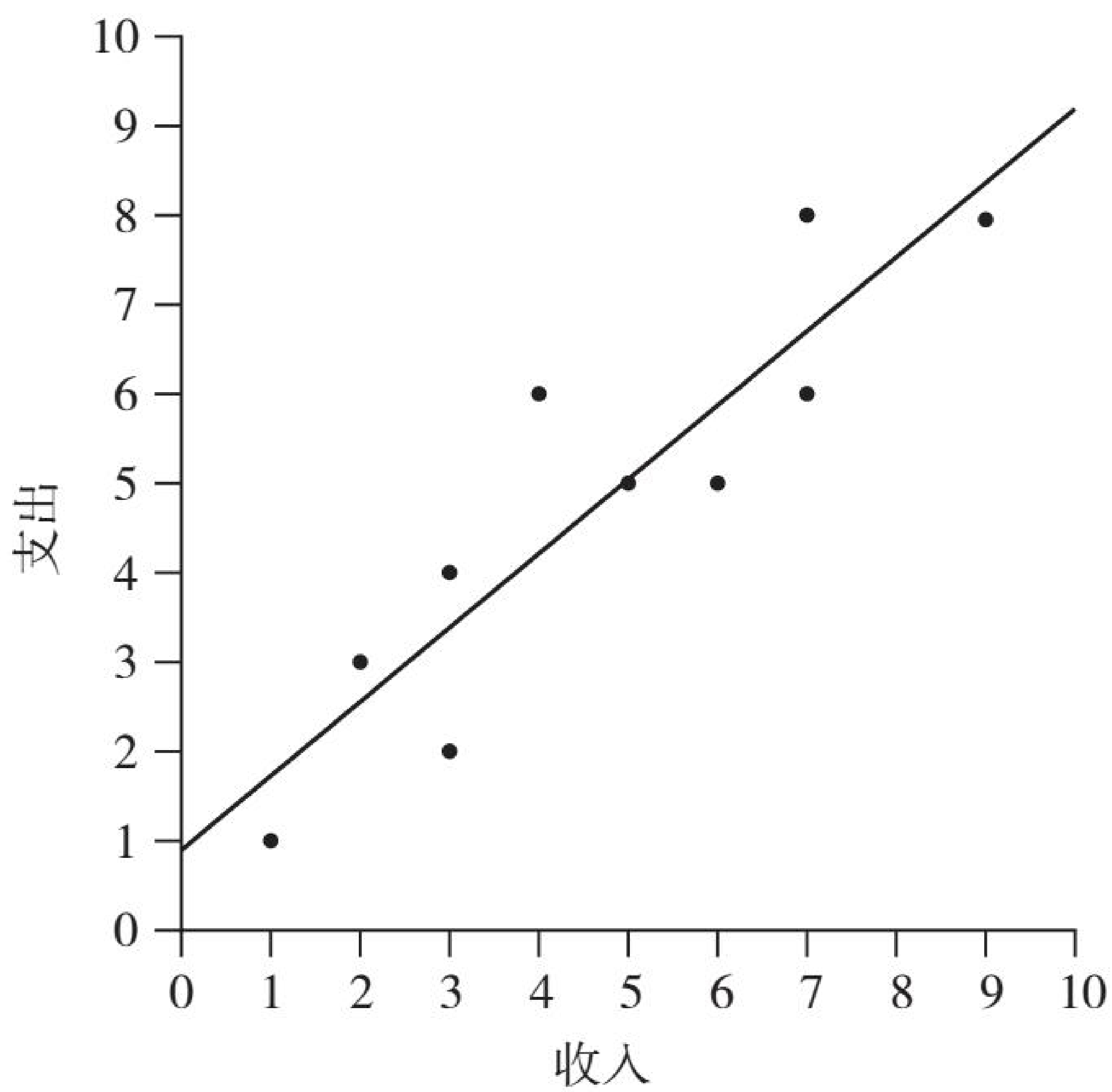


图7.1 收入和支出的正相关系散点图

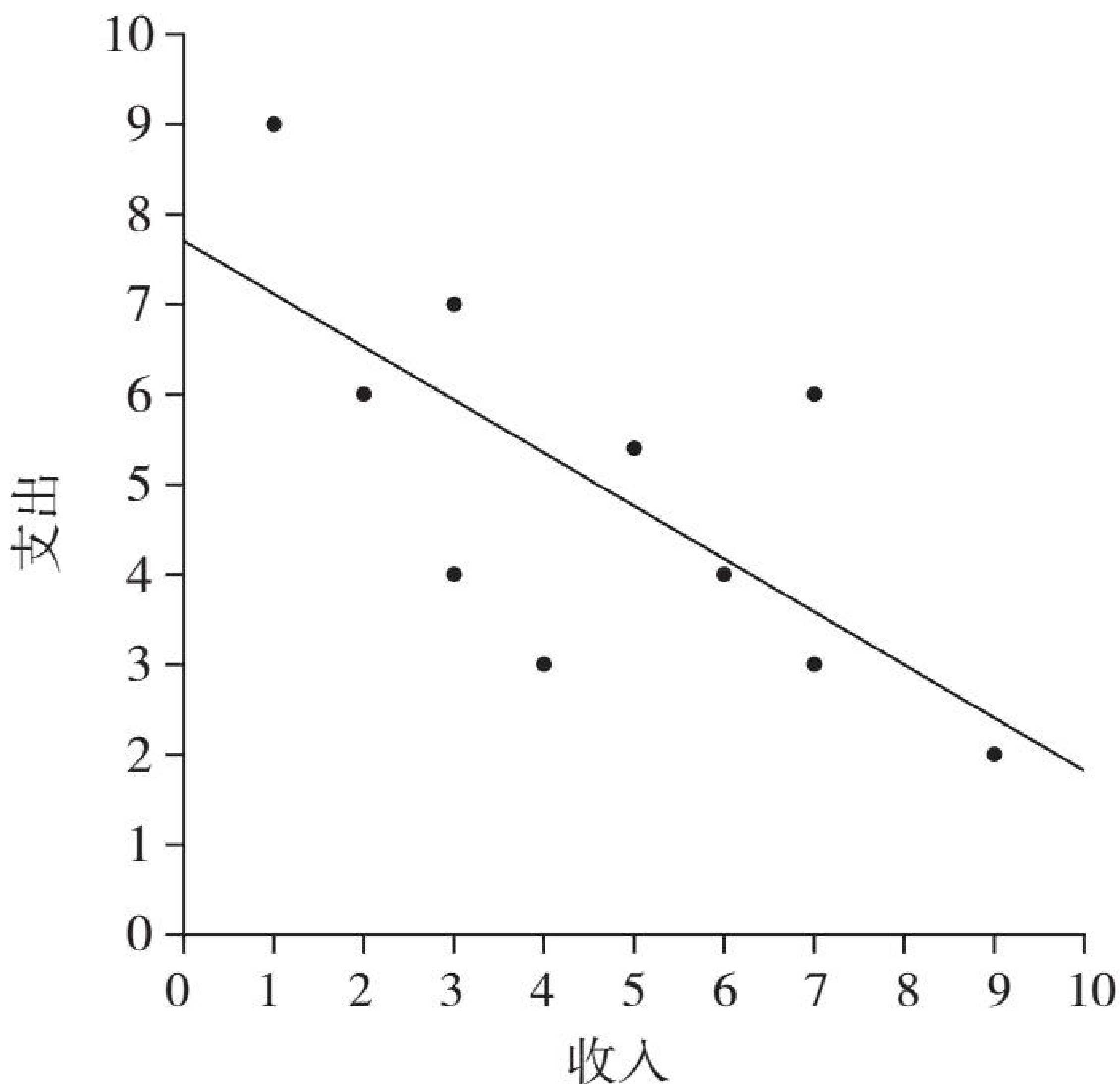


图7.2 收入和支出的负相关系散点图

系数b计算财富状况保持不变时收入增加对支出的影响，而系数c计算收入保持不变时财富增加对支出的影响。推算这些系数的数学过程非常复杂，但是原理很简单：为用来推算模型的数据选择能最佳地预测消费性支出的推算。

我们已经在第4章了解到，在比较支出、收入和财富这些都会随时间推移而增加的变量时会出现“假性相关系数”。为确保不被假性相关系数误导，我要看的是去除通胀因素后的支出、收入和财富的年度百分比变化。

我使用统计学软件来计算美国年度数据的回归线：

$$C = 0.62 + 0.73Y + 0.09W$$

财富保持不变，收入每增加1%，支出预计会增加0.73%。收入保持不变，财富每增加1%，支出预计会增加0.09%。图7.3为实际支出的百分比与预测支出的百分比的变化对比图，相关系数竟惊人地达到0.82。

财富的系数看似很小，但是财富的变化通常很大。有好几年，财富增加或下降的幅度超过10%，根据我们的模型预测，消费性支出会下降0.9%，这就形成了经济扩张和衰退之别。

多元回归模型的效力极大，远比简单的相关系数大得多，因为它将多个解释变量考虑在内。这就是为什么多元回归模型是较重要的统计学工具之一。

然而，多元回归模型用于数据挖掘时也非常容易出现滥用的情况。

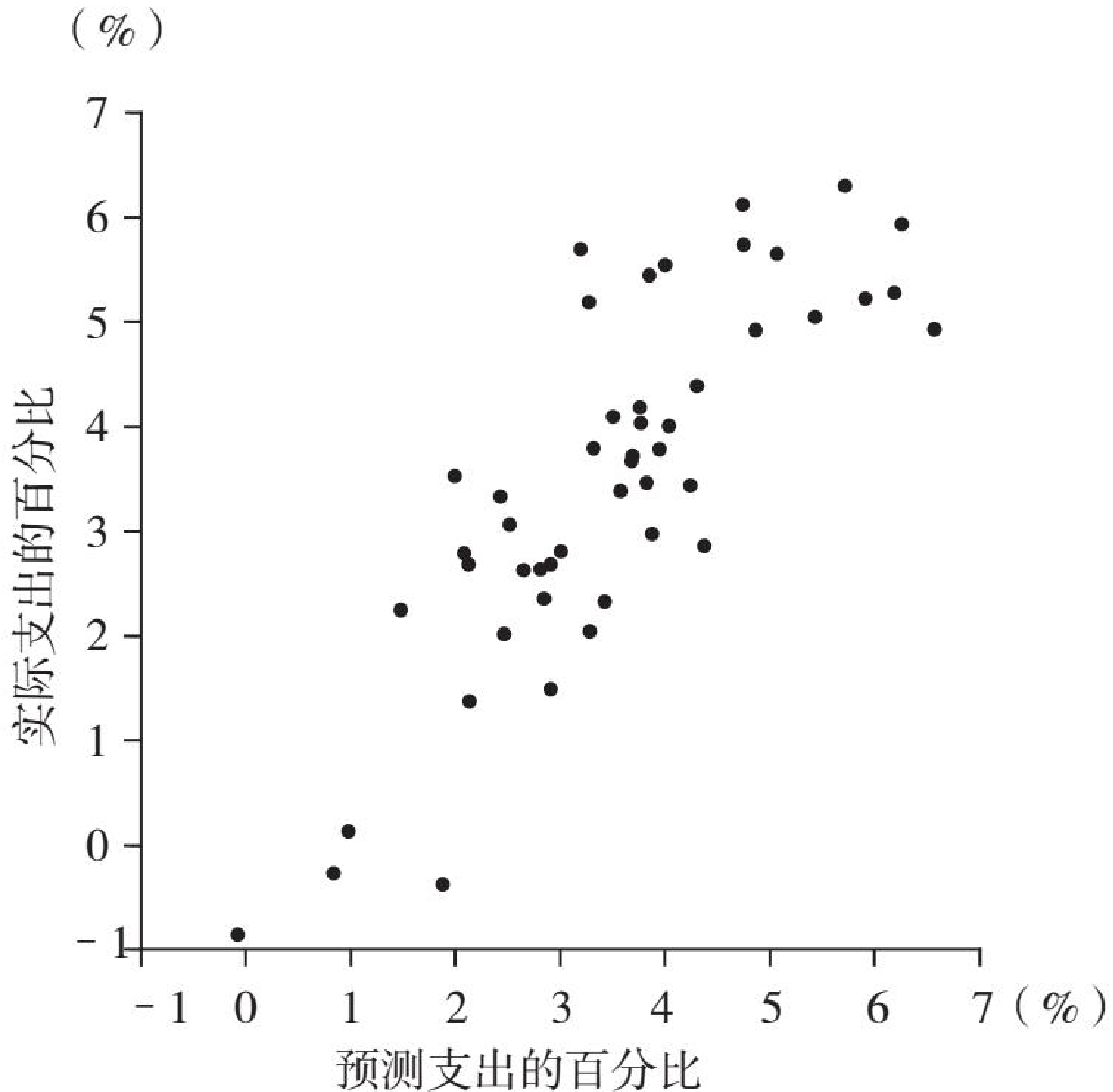


图7.3 美国家庭的预测支出与实际支出的百分比变化

预测总统大选

在统计学课上，我要求学生列出自认为可以决定总统大选结果的因素。他们提到了经济、候选人个性、国家是否处于战时状态等。我将他们的想法写在白板上，然后展示我的模型。

100多年来，美国总统大选通常都是民主党和共和党两党的总统候选人之争。执政党的总统候选人要么是总统本人，要么是总统所在党派的提名候选人。2012年，竞选第二任期的贝拉克·奥巴马就是执政党的总统候选人。2016年，取代已连任两届的奥巴马而参加竞选的希拉里·克林顿，就是总统所在党派的提名候选人。

执政党具备很多优势，包括有更便捷的渠道接触媒体、筹集资金。执政党可以吸引渴望稳定和对现状满意的民众。另外，对经济、战争等问题感到不满的选民可能会投票支持挑战者，即替换候选人。据估计，总统候选人与挑战候选人的优势比为4：6，尽管最终结果显然取决于具体候选人和历史环境。

如果我告诉学生，我只根据执政党的提名候选人是否为总统就能预测执政党在两党投票中的得票率，他们肯定认为我是在开玩笑。他们完全有理由这么认为。我们都知道在任总统什么时候表现好（罗纳德·里根得票率为59%），什么时候表现不好（吉米·卡特得票率为44%）。

但是，如果我还将候选人是否曾任州长和参议员等因素都考虑在内会怎么样？我向学生展示了以下多元回归模型，该模型是我利用过去10次总统大选（1980—2016年）的结果推算而来的：

$$i\% = 78.31-7.35iP-13.07iV+7.93cV-27.20iS+14.75cS-34.46iG+8.20cG-19.54iR+3.49cR$$

这些变量分别代表：

- i% = 执政党候选人获得的主要党派投票百分比
- iP = 执政党候选人是总统时等于1，否则等于0
- iV = 执政党候选人担任过美国副总统时等于1，否则等于0
- cV = 挑战者党候选人担任过美国副总统时等于1，否则等于0
- iS = 执政党候选人担任过美国参议员时等于1，否则等于0
- cS = 挑战者党候选人担任过美国参议员时等于1，否则等于0

iG = 执政党候选人担任过美国州长时等于1，否则等于0

cG = 挑战者党候选人担任过美国州长时等于1，否则等于0

iR = 执政党候选人担任过美国众议员时等于1，否则等于0

cR = 挑战者党候选人担任过美国众议员时等于1，否则等于0

我并不考虑经济、候选人个性以及我的学生认为重要的其他因素。我选择一些依稀相关的因素，并得到了准确无误的关联，因为我的等式可以完美地预测这10次总统大选的所有结果。例如，我的模型对希拉里·克林顿在2016年两党投票中的预测结果为51.11%，正等于她的实际得票率。

当我的学生看到模型与数据完全匹配时，他们不禁认为我已经找到了预测总统大选的神器。我的模型并不包括他们认为重要的任何因素，但是它看上去很合理，因为它使用了与总统候选人背景相关的解释变量。最重要的是，我的模型与数据非常吻合，因此它肯定正确，是学生自己犯错了。

然后，我又给他们展示第二个完全符合1980—2016年10次总统大选数据的模型：

i% = 84.79 - 1.62T1 - 0.30T2 - 0.04T3 - 0.54T4+2.94T5 - 0.39T6 + 0.60T7 + 0.14T8 - 1.05T9

这9个解释变量均为大选之日的最高气温，分别来自9座城市，并且这些城市所在的大州只有极少数选票：

T1 = 蒙大拿州博兹曼市的最高气温

T2 = 内布拉斯加州布罗肯鲍市的最高气温

T3 = 佛蒙特州伯灵顿市的最高气温

T4 = 缅因州卡里布市的最高气温

T5 = 怀俄明州科迪市的最高气温

T6 = 特拉华州多佛市的最高气温

T7 = 西弗吉尼亚州艾尔肯斯市的最高气温

T8 = 北达科他州法戈市的最高气温

T9 = 爱达荷州波卡特洛市的最高气温

之所以选择这些城市，是因为我喜欢它们的名字，也能找到它们早至1940年的每日天气数据。

现在，我的学生都感到困惑了，同时还有很多人心存疑虑。这些都是我一手捏造的吗？博兹曼市或布罗肯鲍市的气温怎么会对总统大选造成实质影响呢？为什么执政党候选人获得的选票与博兹曼市的温暖天气存在负相关系数，而与科迪市的温暖天气存在正相关系数？完全没有符合逻辑的解释，但这个模型却与数据非常吻合。

总统大选可能会受到天气影响。这可能是搜遍数据才能发现的、意料之外的关系。我可能偶然做到了知识发现，证明了数据挖掘的威力。你是否也不禁信以为真了？

所以，我决定让模型看似更加荒唐。我推算出第三个与1980—2016年10次总统大选数据完全相符的模型：

i% = 33.73 - 0.01R1 + 0.26R2 + 0.21R3 + 0.20R4-0.01R5 + 0.19R6 + 0.01R7 - 0.33R8 - 0.18R9

这一次，解释变量的确都是随机得来的。我使用了计算机软件随机生成两位数的数字，这些数字与现实世界没有一点关系，与总统选举年期间美国发生的事情更没有关系。但是，该模型与数据的匹配度还是非常高。

尽管我的学生满腹疑团，但这一切确实并非我无中生有。不过，我确实有个秘诀。

独家秘诀

假设我想解释为什么2016年底的股价会比2015年底的高10%，并且我声称这一切都是因为天气。具体来说，是因为位于加利福尼亚州中央山谷的波特维尔小镇的天气，我父亲就在那里长大。你会认为我疯了，而如果我真的这么想，那么你说得也没错。但是，你先听我把话说完。

图7.4所示的数据散点图为2015年和2016年最后一天的标准普尔500指数以及波特维尔的最低气温。图中显示两者存在绝对完美的相关关系。这两个变量之间的相关性为1。股价完全可以根据我父亲家乡的气温变化来预测。谁能想到呢？

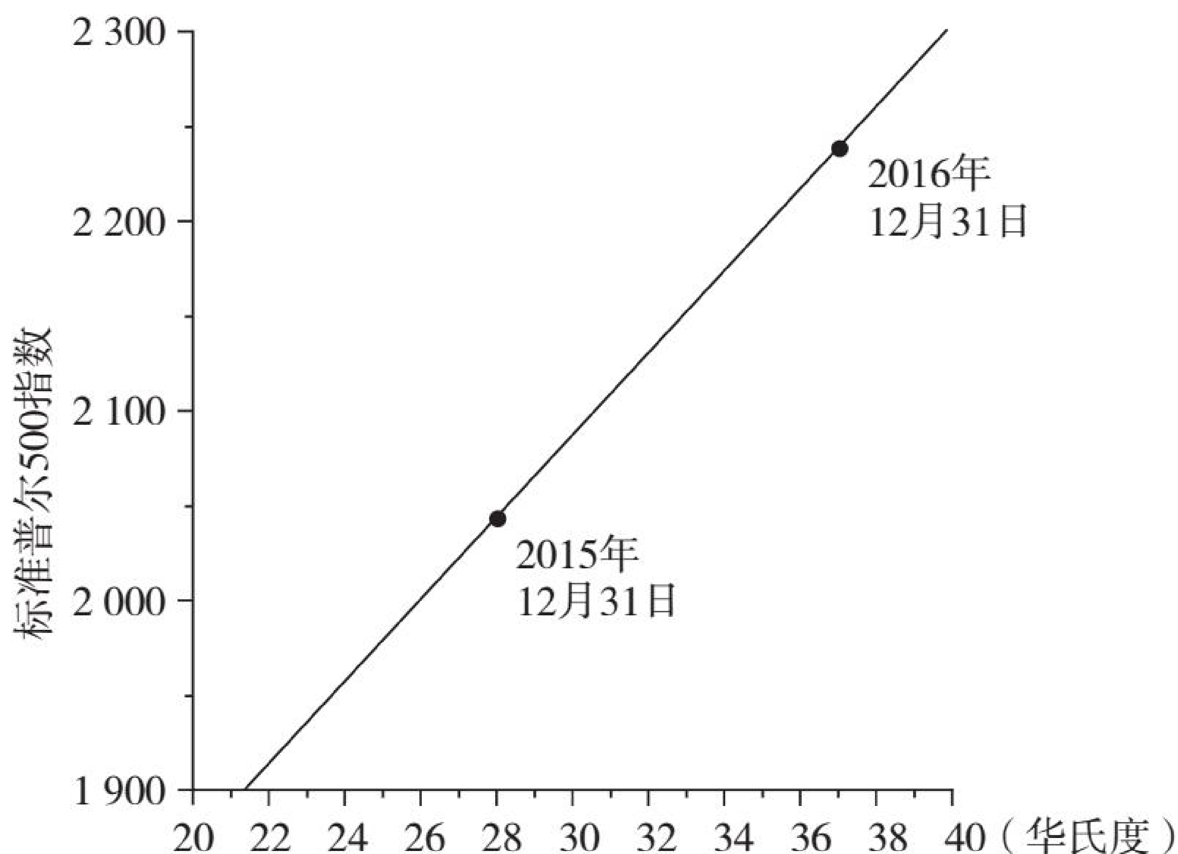


图7.4 用加利福尼亚州波特维尔小镇的最低气温预测股价

秘诀（这当然少不了）就是，散点图中的两点之间总会存在完美的线性关系。我还可以选择1974年和1997年出生的、名叫克莱尔的新生儿数量，或是圣安东尼奥马刺篮球队在2012年和2015年的获胜场数。这

些数据和标准普尔500指数之间同样会存在完美的线性关系，因为连接两点总会出现一条直线。

然而，这种拟合关系却毫无用处。任何试图通过波特维尔的气温来预测股价的人，都会以失败而告终。

我在二维图表中使用两个数据点，说明这种荒唐的想法适用于采用更多数据的、更复杂的模型。图7.4使用一个解释变量（波特维尔的气温）完全匹配了两种观察结果。如果有3种观察结果，两种解释变量也完全匹配。即便有10种观察结果，9种解释变量也是一样的情况。

这就是我得出上述三个预测10次总统大选的模型的方法，一个比一个离谱，秘诀在于使用9个解释变量，仅此而已。这9个解释变量也没有什么特别之处，任何9个都可以。重点在于，我使用这9个变量的目的是预测10次大选。

这就是所谓的“过拟合”（overfitting）数据的极端例子。在任何实证模型中，我都能通过增加越来越多的解释变量，来提高模型的解释力——在极端的例子中，可以将其提高到精确吻合的程度。变量是否合理几乎无足轻重。

这种建模方法也就是常说的“厨房水槽法”，即一股脑把所有解释变量统统塞进模型中。无法避免的问题是，即使模型与原始数据吻合度很高，使用新数据来预测也丝毫不起作用。波特维尔的天气不能准确预测股价，除非“瞎猫碰到死耗子”。我做的包含9个变量的总统大选模型也无法准确预测其他总统大选结果，除非歪打正着。

回看1980年之前的10次总统大选，就能看清我做的总统大选模型的缺点。如图7.5所示，运用时任总统情况和挑战者的数据得出的模型1与1980—2016年间的10次总统大选结果完全吻合，但是与1980年之前的10次总统大选的结果却截然不同。该模型预测理查德·尼克松会在1972年的大选中惨败，普选得票率仅为29%。可实际上他以绝对优势取胜，普选得票率高达62%。尼克松拿下了除马萨诸塞州之外的各个大州，这个“海湾之州”的一些民众还在保险杠上贴了贴纸：“别怪我们。”

模型1对1956年和1964年总统大选的预测结果更是一塌糊涂，竟然预测德怀特·艾森豪威尔在1956年的得票率高达几乎不可能的79%（实际结

果为58%），还预测林登·约翰逊在1964年的得票率低至几乎不可能的26%（实际结果为61%）。

我过度拟合了最近那10年的总统大选数据，随后尝试预测早前的大选结果（以失败而告终）。我同样也可以通过过度拟合早前10年的总统大选结果来推算系数，然后再用这个模型来预测最近10年的总统大选结果。如图7.6所示，修订版模型与1940—1976年之间的10次总统竞选结果完全吻合，但对最近10年的总统大选的预测结果却糟糕透顶。

图7.5 使用1980—2016年过拟合数据预测总统大选

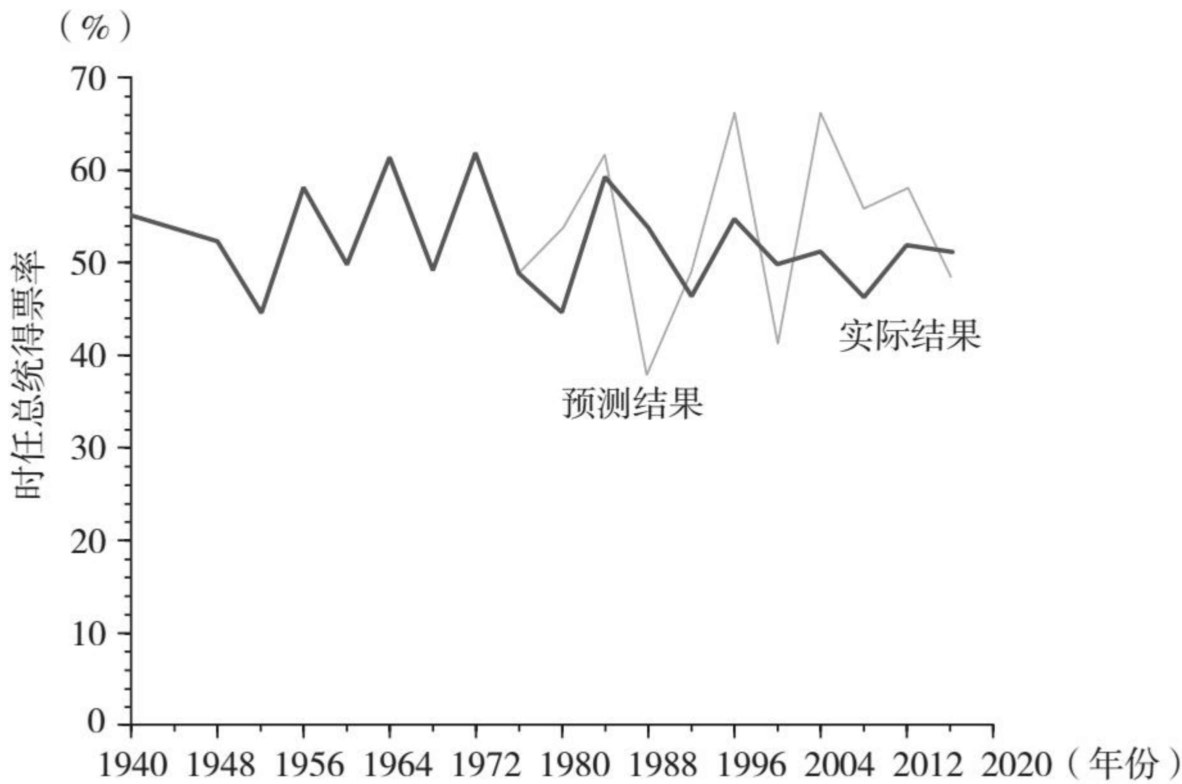


图7.6 使用1940—1976年过拟合数据预测总统大选

模型2和模型3的情况与此如出一辙。如图7.7所示，气温模型与用来推算该模型的数据完全吻合，但是对其他年份大选结果的预测却不尽如

人意。该模型预测富兰克林·罗斯福在1940年大选中的得票率为-11%（没错，就是负数），而他的实际得票率为55%。

坦白说，使用9个解释变量来预测10次总统大选是个极端例子。我这么做是想说明一个普遍原理，那就是即使在回归模型中增加毫无意义的解释变量也会提高模型的吻合度。

在预测总统大选的天气模型中，我们不需要添加全部9个解释变量才能达到很高的吻合度，即便只有5个解释变量（即伯灵顿市、科迪市、多佛市、艾尔肯斯市和波卡特洛市的天气），天气模型预测结果与实际得票率之间的相关系数也会达到0.94：

i% = 72.75 - 0.38T3 + 0.59T5 + 0.40T6 - 0.38T7 - 0.65T9

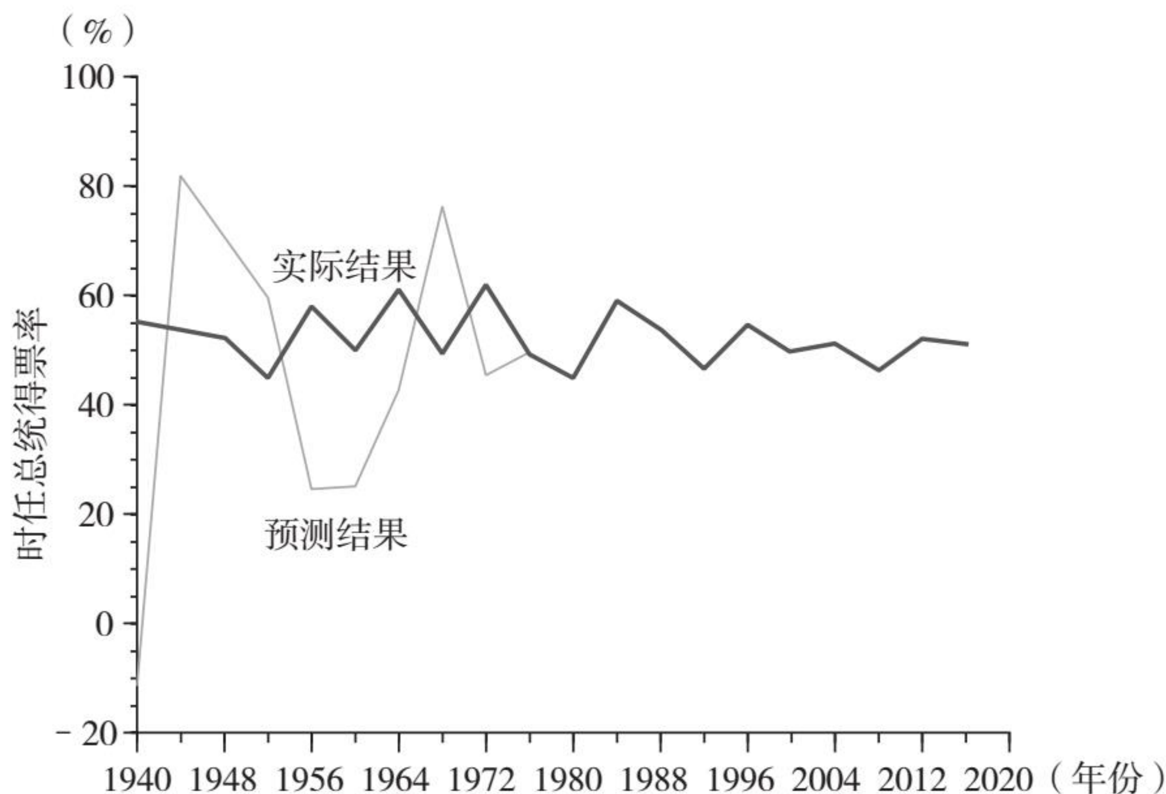


图7.7 使用1980—2016年的天气数据来预测总统大选

如图7.8所示，显然，这个包含5个解释变量的天气模型与1980—2016年的大选数据高度吻合，而与1940—1976年的数据大相径庭：

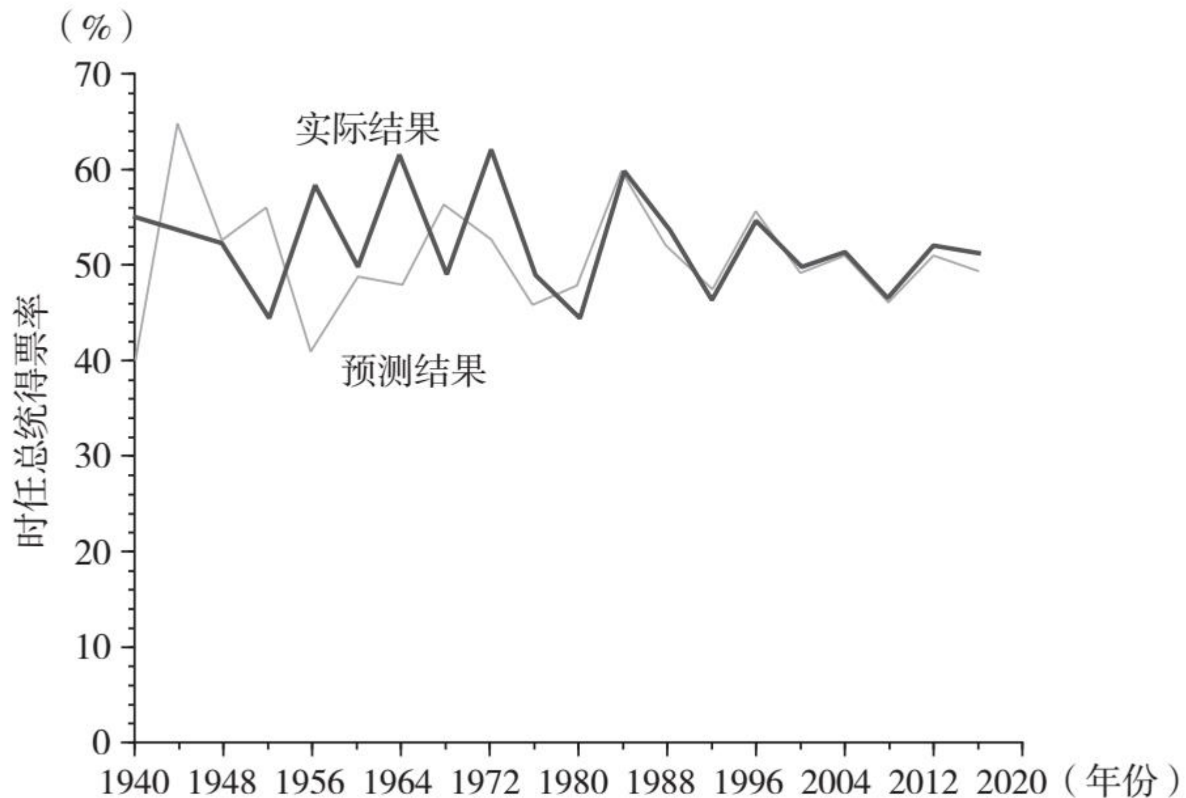


图7.8 使用5座城市的气温来预测总统大选

我们还能以更少的解释变量达到很高的吻合度。包括4座城市（伯灵顿市、科迪市、艾尔肯斯市和波卡特洛市）的天气数据的模型与1989—2016年大选结果的相关系数为0.86，而包括3座城市（科迪市、艾尔肯斯市和波卡特洛市）的数据时相关系数为0.79。

如果将该模型与1940—1976年的数据匹配，情况也是一样。包括4座城市（布罗肯鲍市、多佛市、艾尔肯斯市和法戈市）的天气数据的模型与1940—1976年大选结果的相关系数为0.89；而包括3座城市（布罗肯鲍市、艾尔肯斯市和法戈市）的天气数据时相关系数为0.86。

我选出这些城市的依据是什么呢？我有25座城市每日最高和最低气温的数据，使用数据挖掘软件将这50个变量的所有可能组合统统考虑在内，然后识别出与总统大选结果吻合度最高的组合。

结果显示，1980—2016年吻合度最高的城市与1940—1976年吻合度最高的城市截然不同，因为该模型没有理论基础。这些我用来寻找假性

相关系数的数据本质上是随机的。任何差强人意的数据挖掘程序都能得到同样毫无意义的结果，而且还根本不知道这些都是无稽之谈。

如果解释变量减少，随机变量模式也还是能与数据高度吻合，如解释变量减至5个，相关系数为0.97；减至4个，相关系数为0.95；减至3个，相关系数为0.89。一切都与天气模型非常相似，如果某一年的数据没有拿来推算该模型，那么这个模型对该年度大选结果的预测毫无用处。

由此得出的结论不可忽视。数据挖掘能轻易发现包括多个解释变量的模型，即便解释变量与所要预测的变量毫无关系也能与数据达到惊人的吻合度。不足是，数据挖掘软件不能评估模型是否合理，因为对计算机软件来说，数字只是数字而已。

我们如何分辨所发现的模型是真实还是虚假的呢？只要懂得利用人类对变量的认识，就能判断所发现的模型是否具有逻辑基础。

我一直在强调这一点，是因为我与很多聪明的相关人士都交谈过，他们虽是出于好意，但始终不能完全理解找到偶然性的模型和关联性是多么轻而易举的事情，其中还包括大多数和我交流过的数据挖掘者。很多人都模糊意识到可能存在假性相关系数，但尽管如此，他们还是相信模型和关联性的统计学证据足以证明它们就是真实存在的。

2017年，《华尔街日报》的首席经济评论员格雷格·伊普采访了一家为企业开发人工智能应用程序的公司的合伙创始人。伊普复述了此人的论点：

> 如果在大学学过统计学，你就会知道如何利用输入来预测输出，例如，基于身体指数、胆固醇和吸烟状况来预测死亡率。可以通过添加或取消输入来提高模型的“吻合度”。
>
> 机器学习使用强大的算法和计算机来分析更多的输入。例如，数码图片中的数百万像素，不仅有数字，还有图像和声音。它从变量组合中衍生出更多变量，直至能最准确地回答问题（如“这是一张狗的图片吗”）或者能最圆满地完成任务（如“说服观看者点击本链接”）。

此言差矣！学习统计学的学生在大学里应该学到的是：仅为了提高适合度就添加或取消输入有百害而无一利。机器学习也是如此。搜遍数

字、图像和声音寻求最佳匹配，这是盲目的数据挖掘，考虑的输入越多，所选变量的虚假度就可能越高。

数据挖掘的根本问题在于：它非常擅长找到匹配数据的模型，但对判断模型是否荒唐可笑完全束手无策。统计学相关系数无法替代专业人士的意见。

为现实世界建模的最佳方法是，从具有吸引力的理论学说开始（如“经济状况会影响总统大选”），然后验证模型。合理的模型可对其他数据做出有用的预测，而不是预测用来推算模型的数据。数据挖掘则是反其道而行之，它没有基础理论，因此无法区分合理与荒谬的模型。这就是为什么这些模型对于全新数据的预测结果并不可靠。

非线性模型

除了通过筛选全部解释变量，数据挖掘算法还能通过大量非线性模型来过度拟合数据。

图7.9所示的简单散点图使用了假设数据。图中的三个观察结果都没有在直线（线性模型）上，但还是可以看出其大致走向，如果X和Y之间确实存在因果关系，则可能有助于预测Y值。

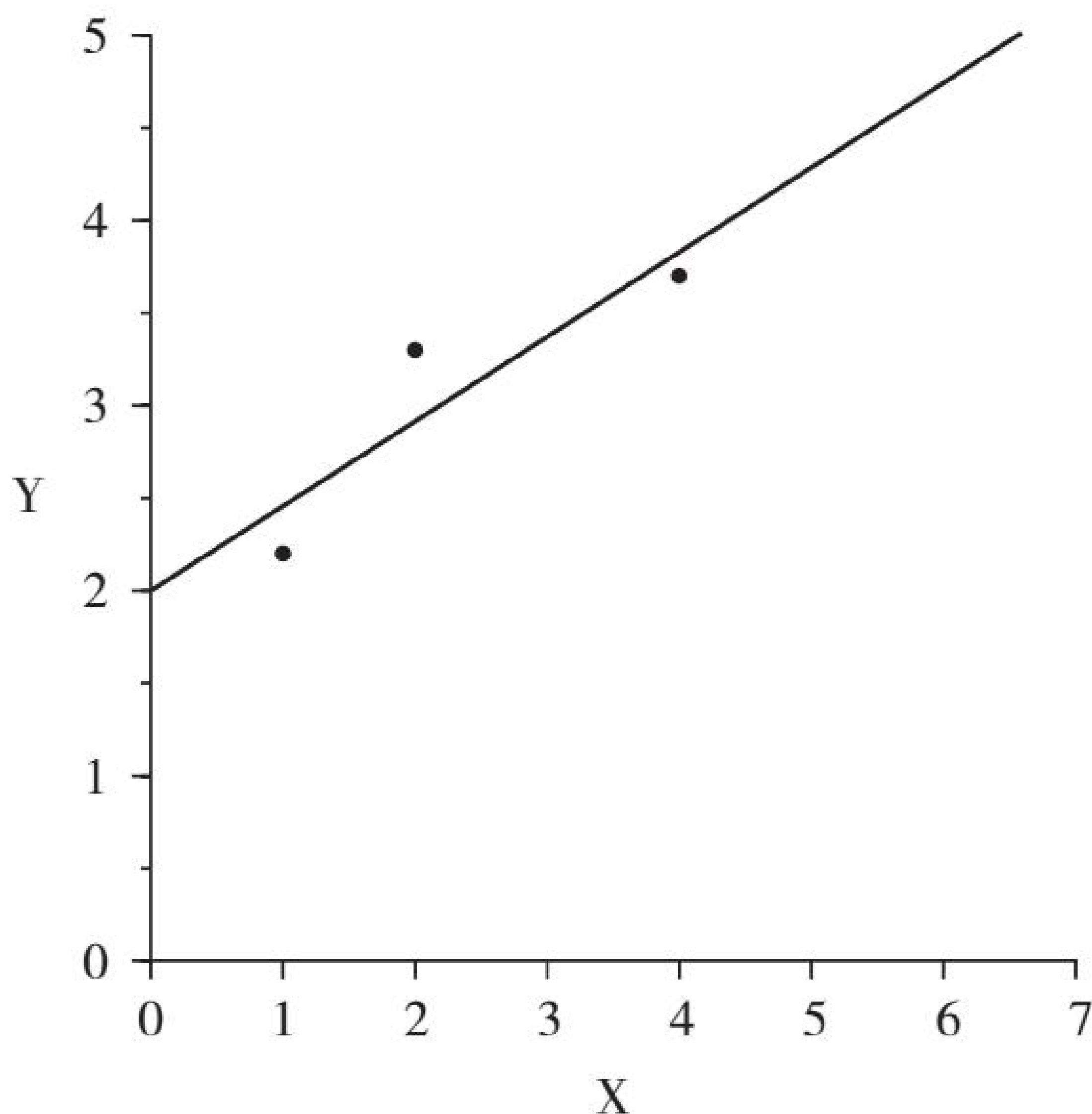


图7.9 线性模型与三种观察结果不吻合

图7.10所示的非线性模型与这三个观察结果完全吻合。可以因此说图7.10的非线性模型是图7.9线性模型的改进版吗？不一定，数据挖掘算法没有合理的方式进行判断。

图7.9的模型显示，X值上升，Y值也上升，增幅保持不变。图7.10的模型显示，X值上升，Y值上升幅度越来越小直至变为负数，X值大于7时，Y值为负数。

要用与模型不吻合的X值来预测Y值，哪个模型更有效呢？看情况。如果X表示家庭收入，Y表示支出，则如图7.9所示，收入增加时，支出也以大致相同的幅度增加，这种说法很合理。但是如图7.10所示，收入增加到某个点时导致支出减少，直至降为负数，这就说不过去了。

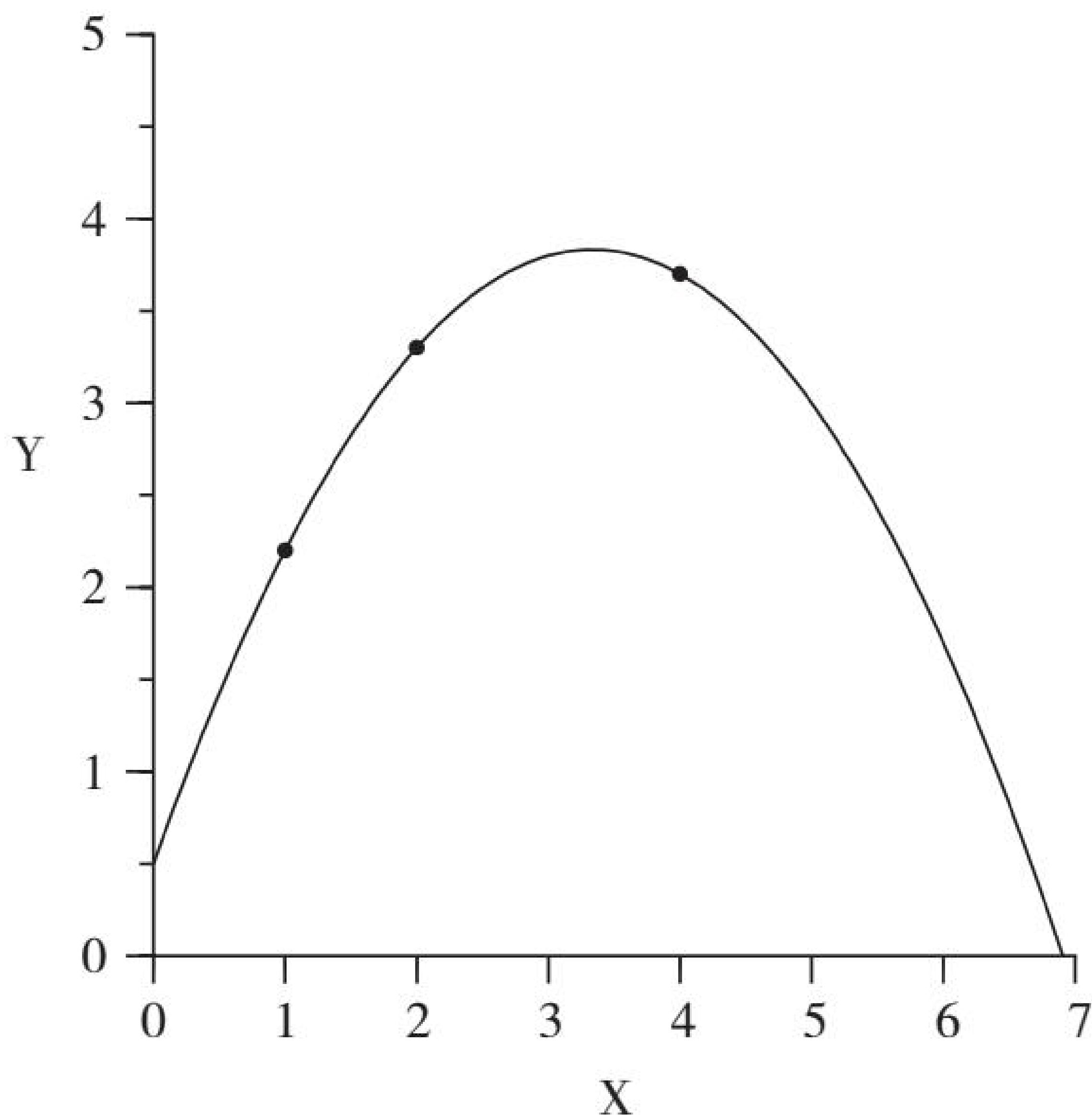


图7.10 非线性模型与三种观察结果完全吻合

此外，假设X表示施给土豆苗的氮素数量，Y表示生长状况。这种情况下，如图7.9所示，即每多施加一点氮素，生长就会快一些，这不合常

理。对比之下图7.10更合理，随着氮素的用量不断增加，其对土豆苗生长的促进作用也会不断减弱。在某一点上，额外的氮素会有碍土豆苗的生长，土豆苗甚至会因为氮素过多而死亡。

数据挖掘算法如何能够决定是图7.9中的线性模型还是图7.10中的非线性模型可以更好地表示建模的事实情况呢？当然不可以只通过看哪个模型与数据更加匹配来决定！我们只能通过专家（即人类）的建议来评估哪个模型更符合现实，才能在这些或其他模型中做出选择。

图7.11展示了更极端的例子。如果存在符合逻辑的解释，这个解释似乎能将直线与所有数据完全匹配，将直线附近的点解释为模型之外其他因素造成的不可避免的波动。除非发生剧烈变动，否则利用线性模型应该可以做出合理精确的预测。

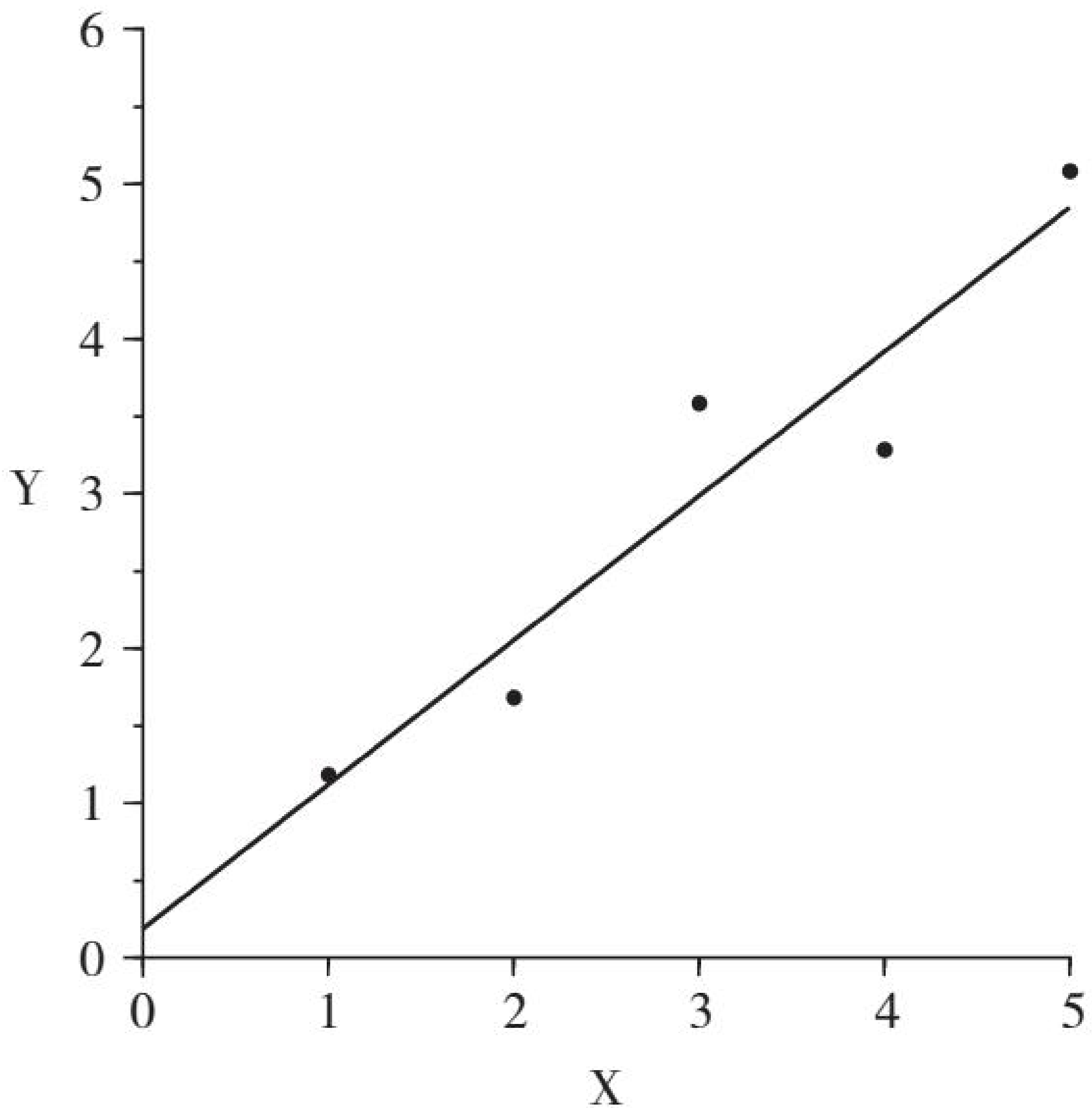


图7.11 合理的线性模型

图7.12显示了数据挖掘算法为了完全匹配数据而选择过度复杂的非线性模型后出现的混乱趋势。尽管与原始数据完全吻合，但只要输入新的X值，该非线性模型的预测结果就肯定会差之千里，甚至会令人匪夷所思。

问题自始至终都在于，数据挖掘算法寻找模型（这也是它非常擅长的事情），但是没有办法评估自己找到的模型。spending（花费）、

income（收入）和wealth（财富）等词语都只是字母组合而已，正如奈杰尔·理查兹用自己不懂的语言玩拼字游戏那样。计算机算法不能分辨模型中应该包括哪些解释变量，也说不出线性和非线性模型哪个更合理。这些都需要人类智慧来做决定。

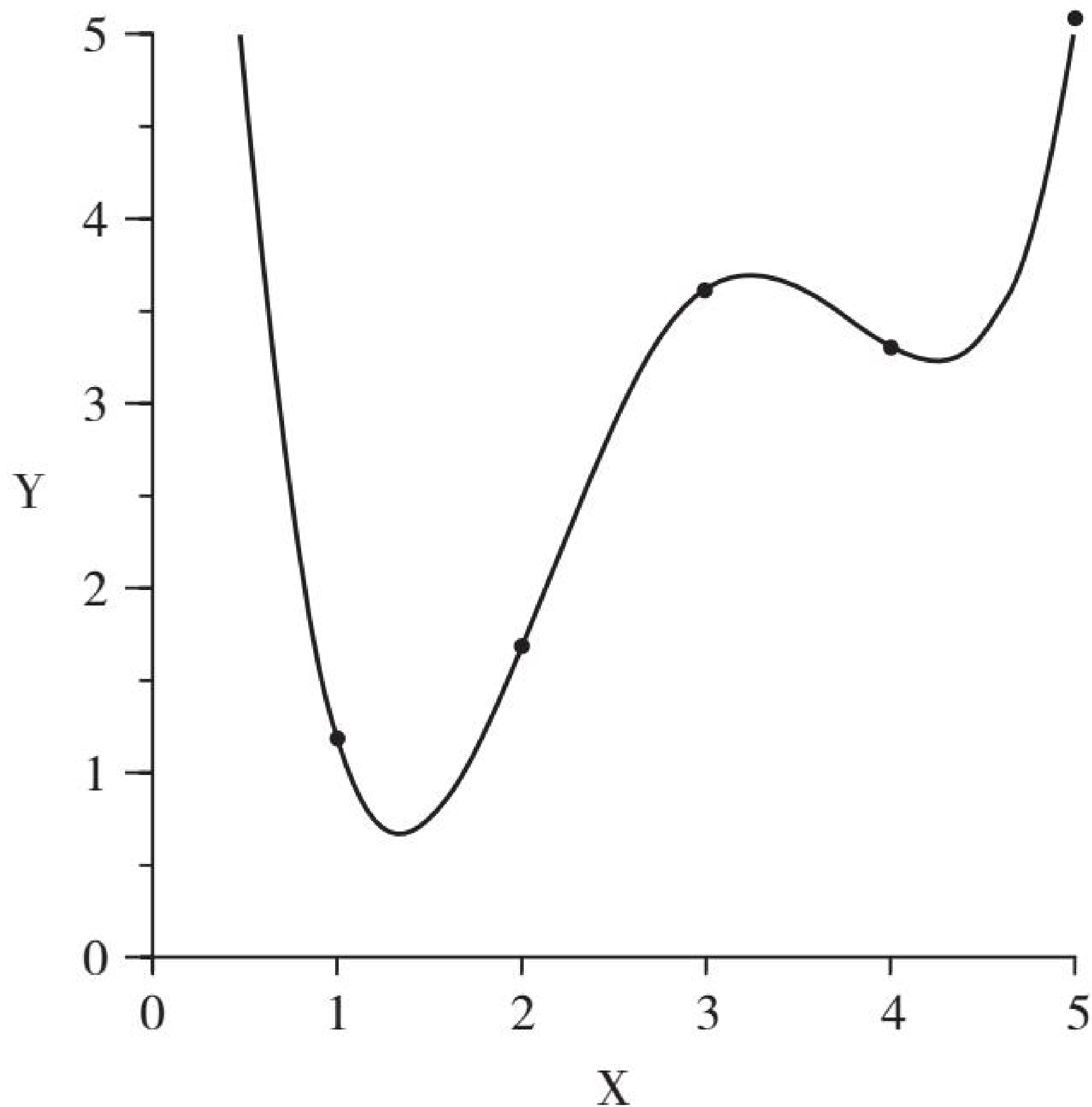


图7.12 不合理的非线性模型



第9章
先吃两片阿司匹林

> IBM的“沃森”在《危险边缘》游戏中夺冠后，得到了铺天盖地的宣传，不过“沃森”的潜在价值更多体现在能够为医生、律师等需要快速准确获得信息的专业人士提供大规模的数码资料库上。

当医生怀疑病人患有某种疾病时，“沃森”可以列出可识别的症状；当医生注意到患者出现异常情况，但不确定这些症状与哪种疾病相关时，“沃森”可以列出可能的疾病；当医生确认患者得了某种疾病时，“沃森”可以列出推荐疗法。在上述每种情况下，“沃森”都会给出多种建议，随附其他相关的可能性，以及它所依据的就医记录和杂志期刊文章的超链接。

“沃森”和其他医学数据库都是宝贵的资源，可以利用计算机的能力来获取、储存和搜索信息。不过，还是有很多地方需要注意。显而易见的一点就是医学数据库远不像《危险边缘》的数据库那么可靠。人工智能算法非常擅长在数据中寻找模型，但它并不擅长评估数据的可靠性和统计学分析的合理性。

如果医生将患者的症状输入黑匣子式的数据挖掘软件并获得建议疗法，但得不到关于诊断或药方的任何解释，就可能导致悲剧性的后果。试想，出现以下情况，你会有何种反应。你的医生说：“我查不出你的病因，但电脑显示要‘服用这些药物’。”或者“我查不出你的病因，但电脑建议动手术”。

任何使用神经网络或数据规约程序的医学软件，如主成分分析和因子分析，都只是勉强能够为诊断和治疗提供解释。病患不知其所以然，医生也不知道，甚至开发黑匣子系统的软件工程师都不知道。总之，没人知道。

“沃森”和类似软件是极佳的参考工具，但它们无法替代医生，因为医学文献通常有误，数据挖掘软件的使用叠加了这些错误。

明早再给我打电话

几年前，我做了一次例行体检，量了身高、体重，回答了两页纸的问题，都是关于我的生活方式的（我不抽烟），还做了一大堆测试。护士量了我的体温、心率和血压，还检测了尿常规和血常规，检测目的具体是什么也不清楚。当天晚上，我接到回馈电话，被告知某项检测（我记不清是哪项了）的结果有些问题。95%的健康人士的该项检测结果都在“正常”范围内，而我的这项检测结果“不正常”，所以显然我的身体是不健康的。

医生说：“不用担心。”她让我吃两片阿司匹林，睡个好觉，第二天再回去复检。我照做了，第二天的复检结果正常，我也松了一口气。

是多亏了那两片阿司匹林，还是前一晚的好觉？可能两者皆非。最有可能的是，这不过为随机噪声。任凭哪个健康人来做那些检测，结果都会出现变动。一天中的不同时段、消化状况和个人情绪都会影响血压。摄取的食物和检测前运动与否都会影响胆固醇的检测结果。设备误差以及读数、记录、解读时的人为失误都容易影响检测结果。

如果一次检测结果碰巧过高或过低，再次检测的结果就可能会接近平均值。这种逆转情况让评估医学疗法的作用变得困难。就我的例子来说，根本不知道是阿司匹林还是睡个好觉起了作用。

有人说：“如果治疗得当，感冒14天就会康复；如果顺其自然，病情也就持续两周。”虽然医生说“明早再给我打电话”时，听上去像是为了少点麻烦，但这就是老方法的大智慧。

即使我感冒之后吃了阿司匹林不见效，第二天早上也还是会有所好转，因为身体有极其惊人的自愈能力。假设你身上有道伤口深到流血了，肌体的血小板会凝固血液，然后结痂修复皮肤。这一切都是身体的自愈，无须任何医学干预。

“明早再给我打电话”的做法可行，原因有二。第一，医学测试无法完全准确检测病患状况。第二，病患的身体能对抗疾病，通常患病之后不进行治疗也都会有所好转。

比起不必要的担心，医学干预的后果更加严重。偶然波动引起的读数异常，会带来不必要的治疗。接受治疗后的检测结果改善，又会不知不觉让人相信是治疗见效了。

假设有一大批人进行体检，其中被检查出胆固醇指标最高的人会被告知要特别注意饮食。我们能预见到他的胆固醇指标会有所改善，即便饮食调节的指导无非就是“吃前请三思”。

此外，我们都知道，止痛药的效果因人而异，大多数医学治疗都是如此，没有完全有效或无效的疗法。如果有效果不显著或因患者情况不同而各异的情况出现，医学测试的结果就取决于哪些人被随机分配到了服用药物的实验组，哪些人被分配到了服用安慰剂的控制组。

统计学家尝试解释上述的随机变化，他们假设差异纯属偶然，然后评估实验组和控制组之间的差异和观察结果一样显著的可能性有多大。

P值小于等于0.05则具有统计学意义。这意味着，没有价值的被测疗法只有5%的机会显示其统计学意义，也就表示仍有5%的无价值疗法会得到具有统计学意义的结果。

医学研究是个弱肉强食的领域，才智过人和竞争力强的科学家一辈子都在为名誉和经费而奋斗，以维持其职业发展。为了达到这一目的，这些科学家需要获得并发表具有统计学意义的结果——必要时不择手段，其中就包括得州神枪手谬误1和谬误2。

研究人员只要通过大量的疗法测试就能得到有统计学意义的结果，即便他们受到了误导，测试的只是无用的疗法，在上百次无用疗法测试后，他们还是会发现其中5%具有统计学意义——这足以促成其文章发表，使经费提案获批。

同样，医药公司能够从临床“验证”有效的疗法中获得巨额利润。确保某些疗法得到支持的一种方法是，测试数以千计的疗法，无论遇到多少统计学障碍，运气都能确保某些无用疗法跨越所有障碍。

下面让我们一起来看三个“得州神枪手”的例子。

我要再喝一杯咖啡

20世纪80年代早期，据全世界顶尖的医学期刊《新英格兰医学期刊》报道，广受赞誉的研究者、哈佛公共卫生学院院长布莱恩·迈克马宏所带领的团队发现“饮用咖啡与胰腺癌有极大关联”。这个来自哈佛大学的团队建议人们不要再喝咖啡，以降低患胰腺癌的风险。在此项研究之前，迈克马宏自己每天都喝三杯咖啡，在此之后他就再也不喝了。

这就出现了得州神枪手谬误1中的问题。该研究旨在调查喝酒或抽烟与患胰腺癌之间的联系，迈克马宏研究过酒类、香烟、雪茄、烟斗，没有任何发现，于是他就继续找，又研究了茶叶。最后，他终于在咖啡上有了发现：胰腺癌患者喝的咖啡多。

如果上述六项测试都单独进行，每项测试都包含一些与胰腺癌无关的因素，那么有26%的概率会在至少一项测试中产生一个具有统计学意义（P值为0.05）的关联，也就是说有26%的机会可以无中生有。

迈克马宏的研究还有另一个缺陷。他将患胰腺癌的住院病人与患其他疾病的病人进行对比，并且这些病人都由同一批医生负责。问题在于，这些医生通常都是胃肠专科医生，他们的很多患者都因为害怕溃疡恶化而戒了咖啡。但胰腺癌患者没有停止喝咖啡，他们中喝咖啡的人更多。所以并非喝咖啡导致了胰腺癌，而是患其他疾病的病人不再喝咖啡了。

后续研究——其中一项来自迈克马宏的团队——也未能证实最初的研究结果。这一次，他们得出的结论是：“据观察，与早前研究相比，喝咖啡对男性或女性都不存在危险。”美国癌症协会也认为：“最近的科学研究表明，喝咖啡和患胰腺癌、乳腺癌等癌症没有任何关系。”

更近期的研究不仅驳斥了迈克马宏最初的研究结果，而且结果显示喝咖啡（至少对男性来说）反而会降低患胰腺癌的概率！

远程治疗

20世纪90年代，年轻的伊丽莎白·塔尔格医生研究了遥远的祈祷和其他积极意念是否能治愈晚期艾滋病患者。40名艾滋病患者被分成两组。祈祷组患者的照片会被发送给有经验的远程治疗师（从佛教、基

督教、犹太教信徒到萨满巫师都有），他们与病患平均相隔约1 491英里。非祈祷组的20名患者则完全靠自己。

此次测试采用“双盲”（double-blinded）程序，塔尔格和病患都不知道哪些病患是祈祷组的，以免影响测试结果。

为期六个月的研究发现，祈祷组的患者就医时间更短，罹患与艾滋病有关的疾病更少。这次研究的结果具有统计学意义，发表在享誉盛名的医学期刊上。人们出于各自的目的引用塔尔格的研究来证明上帝的存在，或是指出传统观念对心智、身体、时间和空间的认识不足。

美国国家卫生研究所给塔尔格拨款150万美元，用以更大规模的艾滋病患者研究和对远程治疗师能否缩小脑癌患者的恶性肿瘤的调查。就在获得拨款不久后，塔尔格自己也被诊断出得了脑癌，尽管世界各地都有治疗师为她祈祷和发送治疗能量，但她还是在四个月后去世了。

塔尔格去世后，其早前对40名艾滋病患者开展的研究也被查出了问题。之前，她计划对比祈祷组和非祈祷组的死亡率，然而，在为期六个月的研究进行了一个月后，“三联鸡尾酒疗法”（triple-cocktail therapy）开始流行，40名患者中只有1人死亡，这表明该疗法的有效性，但它也消除了祈祷组和非祈祷组进行统计学对比的可能性。

于是，塔尔格及其同事弗雷德·西歇尔转而寻找两组之间的其他差异。他们参考了各种身体症状、生活质量测量、情绪评分和CP4+指数，两组患者在这些方面均无差异。塔尔格的父亲曾经试图通过实验证明人类拥有可感知看不见的物体、读心和仅靠意念移动物体的超自然能力。他要女儿塔尔格继续寻找，只要怀有信念，相反的证据就无足轻重，只要继续在数据中搜寻支持自己信念的证据即可。最终，塔尔格找到了——住院时长和医生探访，尽管医疗保险肯定会使问题雪上加霜。

随后，塔尔格和西歇尔读到了一篇列举了23种与艾滋病相关的疾病的文章。他们或许可以寻找两组实验对象在这23种疾病上的差异。不幸的是，由于采取“双盲”安排，这些疾病的数据均未被记录。塔尔格和西歇尔坚持不懈地仔细研读受试对象的医疗记录，即便他们现在已经知道每名患者的分组情况。完成研读后，他们报告称祈祷组在某些疾病方面比非祈祷组的境况更好。这种积极主动的数据挖掘似乎没有利用数据挖掘软件就完成了。

他们发表的论文显示，该项研究是为调查具有统计学意义的几种疾病而设计的（即得州神枪手谬误1），他们做过的其他测试都未公布，也没有说明最终数据是在研究结束后搜集到的，而且“双盲”控制也被撤销了。他们得到了想要的结果，或许是因为他们的坚持，或许是因为数据不再是双盲状态。

塔尔格的美国国家卫生研究所的研究在她去世后仍在继续。祈祷组和非祈祷组在死亡率、患病或症状方面都未发现有意义的差异。另一项规模更大的研究由哈佛医学院的研究人员执行，观察了1 800名处于冠状动脉搭桥术后康复期的患者，还是没有在祈祷组和非祈祷组的患者间发现明显差异。

癌症群

20世纪70年代，流行病学家南希·韦特海默和物理学家埃德·利珀驾车穿过科罗拉多州丹佛市去考察一些人的住所，这些人未满19岁便因身患癌症离开了人世。他们试图发现这些人住所的共同特征。两人注意到，很多罹患癌症的人都住在大功率电力线附近，因此得出结论：暴露于电力线的电磁场中会导致罹患癌症。

记者保罗·布罗德为《纽约客》写了三篇文章，报道了关于电力线和癌症相关系数的其他奇闻逸事。他还做出了不详警告：“数以千计没有戒备的儿童和成人会罹患癌症，其中很多人都会英年早逝，他们本不该遭此厄运，一切只因他们暴露在电力线的电磁场中。”

这种言论随之在全国造成轰动，为咨询专家、研究人员、律师和包括高斯计（测量磁感应强度的仪器）在内的各种装置提供了有利可图的机会，人们可以用高斯计在家测量电磁场的强度（电磁场读数高的房间会被封住，只用作储物间）。幸运的是，政府并没有扯掉整个国家的电力线。

此次恐慌事件的问题在于，即使癌症患者在人口中只是随机分布的，数据挖掘都更有可能发现受害者在地理上集中的地方。为了说明这一点，我虚构出一个有1万名居民的城市，其住所均匀分布于整座城市，每个人患癌的概率都是1%（我忽略了家人一起居住的情况和年龄因素）。然后，我使用计算机随机数字生成器来决定谁是这座虚构城市

中的癌症患者。据此得出的癌症患者分布如图9.1所示。每个小黑点代表住着一名癌症患者的一户人家，而白色区域即无癌症患者居住。

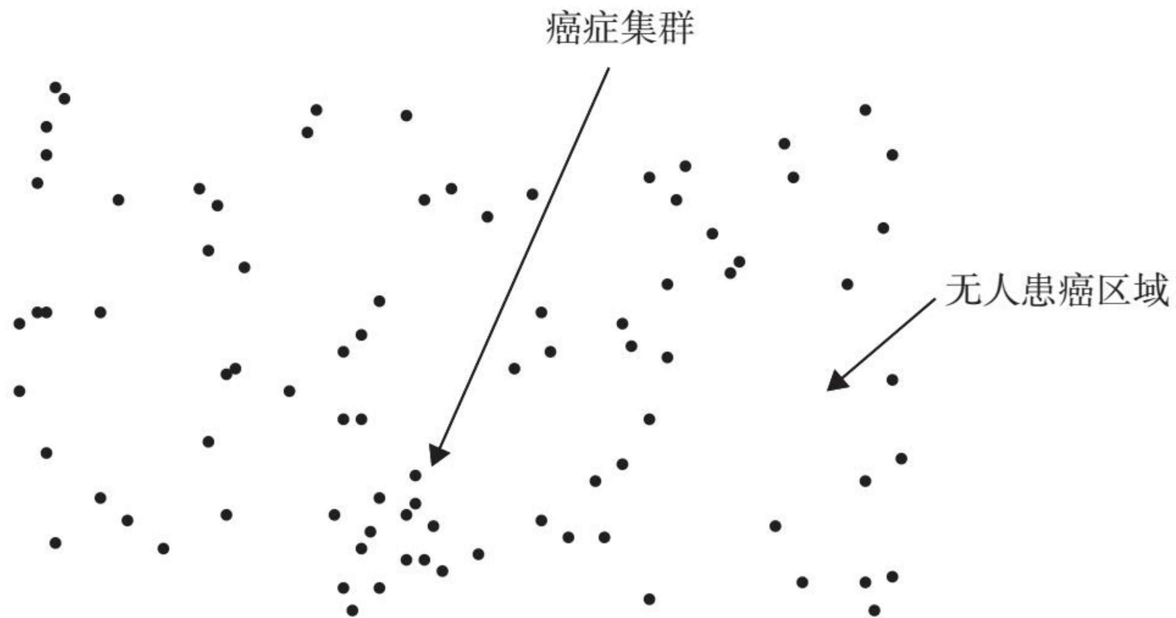


图9.1 癌症患者分布图

随便一个像样的数据挖掘软件都能轻易发现，图9.1的底部明显有一处癌症患者集中地。如果这座城市真实存在，我们就可以驾车到患者住所附近，肯定能得到一些特别发现。或者使用数据挖掘软件搜遍数据，寻找异常状况。如果我们再将住在球场附近的居民患癌率与住所远离球场的居民患癌率相比，猜猜结果如何？球场附近的患癌率更高，这表明住在球场附近可致癌。

图9.1还显示了癌症堡垒，即无人患癌的区域。利用数据挖掘软件或驾车到附近瞧一瞧，一定会得到一些关于这个无人患癌区域的特殊发现。可能该地附近建有水塔。如果我们再将住在水塔附近的居民患癌率与住所远离水塔的居民患癌率相比，一定能发现水塔附近的患癌率更低。这就是我们选择这个区域的原因——这里没人患癌。

无论是在球场还是水塔附近，都存在同样的问题——得州神枪手谬误2。如果我们使用数据来创造理论（小联盟球场会致癌，水塔可防癌），数据当然会支持理论了！怎么会有相反结果呢？我们会捏造出与数据不吻合的理论吗？

用来创建理论的数据肯定不适于再来检验该理论。我们需要全新的数据。其他国家的研究没有发现电磁场和癌症之间存在关联。以啮齿动物为对象的实验研究发现，比电力线所产生的更强的电磁场对死亡率、患癌率、免疫系统、生育率或出生缺陷率都没有影响。

对电力线的恐慌有什么理论基础吗？科学家非常了解电磁场，并没有任何合理理论能证明电力线的电磁场会致癌。电力线的电磁能量远比月光的电磁能量弱得多，其电磁场也比地球的磁场更弱。

权衡理论论证和实验结果后，美国国家科学院得出的结论是：电力线并没有造成公共健康危险，无须提供经费开展进一步研究，更别说撤掉电力线了。全美顶尖医学期刊也发声力挺，同意不应再把研究资源浪费在这个问题上。

1999年，《纽约客》发表了一篇题为“癌症集群之谎言”（The Cancer-Cluster Myth）的文章，含蓄地驳斥保罗·布罗德早先的报道。尽管如此，癌症集群具有意义的想法还是继续存在。互联网上，由政府赞助的交互式地图可按地理区域显示各种癌症的发病率，精细到人口普查的街区。每年都需要花费数百万美元来维护地图数据，虽然数据是最新的，但很可能具有误导性。其中一个交互式网站拥有22种癌症、2种性别、4个年龄段组别、5个种族和3 000多个县的癌症死亡率数据。从数百万种可能的相关系数中，数据挖掘软件一定可以轻易发现令人恐惧的相关系数。

为了缓解这种恐惧，美国疾病控制与预防中心创建了网页平台，任何人都可以在此报告自己发现的癌症集群。即使该中心提醒：“我们会对此进行后续调查，但需要花费多年时间才能完成，结果通常也不能得出定论（也就是说，通常都无法找到原因）。”每年仍有1 000多例癌症集群被举报和调查。

最有理有据的疗法失效了

大量已发表的医学研究都会犯那两个得州神枪手谬误：数据的随机变化只在人们忽略以下情况时有意义，即这些侥幸发现都是靠测试大量理论，或创造理论来匹配数据中的偶然模型才能得到，报告的结果随后便消失得无踪无影。这种模型在医学研究中太常见了，以至于还有专门的叫法——递减效应（decline effect）。

有些研究人员亲眼见过自己的研究出现递减效应，他们都迷惑不解，因此开始白费力气地寻求解释，尽管原因就近在眼前。如果最初的正相关发现皆因得州神枪手谬误，那么随后的结果通常都令人失望也就不足为奇了。这就好比基于偏远城市气温进行总统大选预测那样。

看似有效的无价值疗法只是假阳性结果。另外还有假阴性结果，即有效疗法并未显示出统计学意义。仔细想想，一个测试有5%的机会呈假阳性，就意味着一项经受严格测试的无效疗法，其实验组和控制组之间出现统计学差异的机会为5%。假设假阴性的概率为10%，就表示有效疗法在测试顺利的情况下，无法显示出统计学意义的概率为10%。

如果假阳性的概率为5%，假阴性的为10%，似乎我们每次都应该能分辨出有效和无效疗法之间的区别。实则不然。那要看有多少受试疗法有效，有多少无效。若所有受试疗法中，1%为有效，99%为无效，则结果如表9.1所示。

表9.1 所有经验证的疗法中，有85%为无效

	有统计学意义	无统计学意义	总计
有效疗法	90	10	100
无效疗法	495	9 405	9 900
总　计	585	9 915	10 000

测试10 000种疗法，其中100种有效。这100种有效疗法中，90种会呈现具有统计学意义的结果；而另外9 900种无效疗法中，会有495种呈现具有统计学意义的假阳性结果。因此，共计585种测试具有统计学意义，但其中只有90种为真正有效的疗法，有85%“经验证”有效的疗法实际上毫无价值，这让人难以置信。

这一矛盾反映出有关逆概率的常见困惑。超级联赛的所有运动员都是男性，但所有男性中，只有很小一部分人为超级联赛的运动员。同理，所有有效疗法中，90%都具有统计学意义，但所有具有统计学意义的疗法中，只有15%有效。

任职于希腊约阿尼纳大学、马萨诸塞州塔夫茨大学医学院和加州斯坦福大学医学院的约翰·约安尼季斯以此类运算为依据，发表了一篇以“为何大多数已发表的研究成果都有误”（Why Most Published Research Findings Are False）为题的引起争议的文章。

约安尼季斯在整个职业生涯中都在提醒医生和普通民众，不要轻易相信复制结果无法令人信服的医学测试。他那篇题目惊人的著名文章就采用了我们上述的数学算法，他的假设观点比我们的更加令人确信，而概率的表现也更加糟糕。

除了这些理论性计算，约安尼季斯还汇编列举了在现实世界中“经验证”的疗法最后无效的例子。他在一项研究中检查了45个发表于1990—2003年且广受赞誉的医学研究成果，其中仅有34个能使用更大样本对原始结果进行复制，这其中又只有20个（即59%）证实了最初的结果，7个所述疗法的疗效比最初推算的小得多，剩下那7个疗法则根本一点效果都没有。总的来说，45项研究中仅20项可经证实，这些可都是最享誉盛名的研究啊！对于发表在级别较低的期刊的数千篇研究来说，情况肯定更糟糕。约安尼季斯粗略估算，90%已发表的医学研究成果均有漏洞，其宣称有效的疗法被夸大了效果，有的疗法则毫无效果，甚至更糟。

疾病诊断和治疗中的数据挖掘

传统的统计学测试假定研究人员会以定义好的理论为起始，然后收集合适的数据来验证他们的理论。数据挖掘则另辟蹊径——数据为先，理论在后。因此，可以随意检测所有你想要检测的理论，无论这些理论是否合理。

如果医学疗法没有对整个样本显示出统计学意义，再看看其是否适用于子集；将性别、种族和年龄分开，尝试不同的年龄段；如果该疗法对你最初研究的疾病不起作用，再看看它是否有其他益处。

测试数百种疗法便是得州神枪手谬误1的例子：瞄准数百个目标，只报告那些击中的情况。其他医学研究则有关得州神枪手谬误2：找到一个模式，然后为其编造解释。疾病诊断或治疗都会出现上述情况。

首先讨论一下疾病诊断。假设我们知道100个患者患了某种疾病，不知道另外100个患者患了什么疾病，然后记录下每个人的1 000种特征，比如血液检测、基因信息、种族、发色、瞳孔颜色和住处等。如果我们现在使用数据挖掘软件来彻查这一数据库，肯定会找到一些特征，这些特征在患病人士中比在健康人士中更加常见，而且明显能够很好地预测疾病。

例如，我能够获取87名女性心脏收缩血压读数的数据库，还有每名患者的40种特征的完整信息，有些以数字表示（如年龄），有些按类别区分（如某人是否有吸烟史）。

我使用了数据挖掘软件，来看看根据这40种特征预测血压的结果如何。如果我的模型契合度高，就可以用来识别其他有高血压风险的女性。我们还可以识别出其他高风险因素（可能是吸烟）并建议血压值高的女性改变行为，以降低血压。

该模型非常成功，实际血压和预测血压的相关系数达到惊人的0.72，即23名受试女性的预测心脏收缩血压高于130，符合其中17人的实际情况。图9.2为87名女性的预测值和实际值。

我们还可以只用病患的5个特征（特征1、12、18、23和34）得到预测血压值和实际血压值的相关系数为0.47，这一结果相当不错。因此，医生会重点关注这五个特征，可以预测，甚至可能控制血压升高。

那么这五个特征是什么？随机数字。我捏造了87名女性，使用87这个数字是为了提高研究的真实性。对于其中20个特征，我使用电脑抛硬币的方式来赋予其1值或0值。同样，抽烟者被赋予1值，不抽烟的人则被赋予0值。对于另外20个特征，我用电脑生成了正态分布的随机变量，均值为100，标准差为10。虚假的血压数值也是正态分布，均值为125，标准差为10。我捏造的每一个女性及其每一个特征都与其他女性的捏造特征相互独立，与该女性的虚假血压及其另外39个捏造特征也相互独立。

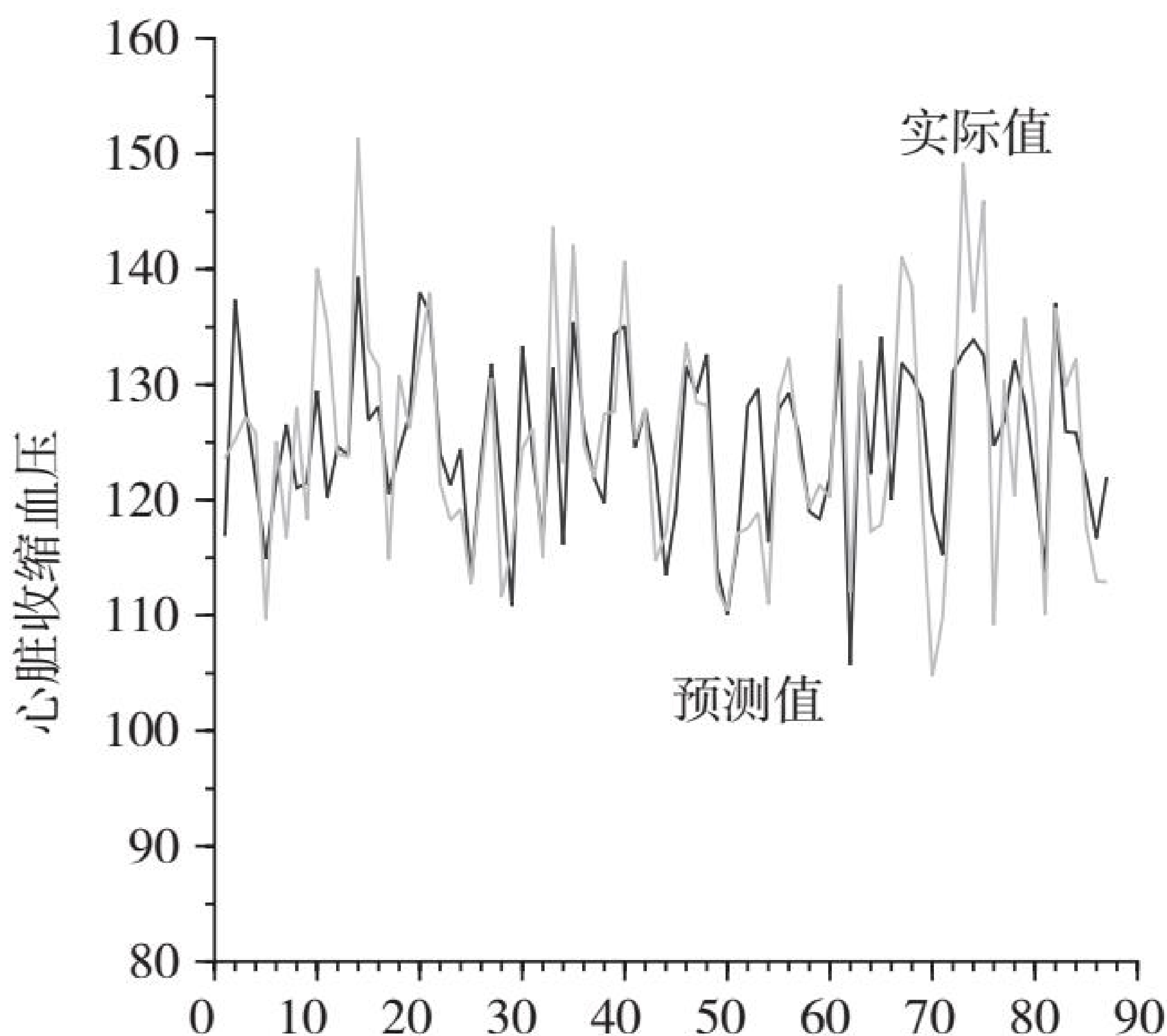


图9.2 心脏收缩血压的预测值和实际值

我在这个虚假的医学数据库中填入随机数字来证明我的观点。即使数据库中记录的特征与所分析的疾病是否存在没有任何关系，数据挖掘软件也会发现具有统计学意义的关系，让人误以为获得了什么有用的发现。

疾病治疗也是如此。假设根据各种各样的医疗状况对病患施行尝试性疗法，使用数据挖掘软件来识别那些有所改善的疾病或疾病组合。即使患者状况的波动完全随机，与他们是否接受了治疗一点关系都没有，还是很有可能存在具有统计学意义的模式表明该疗法对某些状况有效。前面说到的远程治疗研究就是这种谬误的很好例证。

糟糠过多，精粹不足

很多神奇疗法（如胰岛素和天花疫苗）都被医疗研究发现并且证实为有效。然而，很多已发表的研究都有缺陷，这通常是因为那些数据都是为了发表而搜刮得来的。

对“沃森”等医疗建议软件来说，这是无法逾越的难题。它们都非常擅长收集、储存和搜索医疗数据和期刊文章，这一点肯定优于人类。但是它们没有常识或智慧，不知道数字和词语的意思，无法评估数据库中内容的相关性和有效性。它们也无法分辨好数据和坏数据，不能识别哪些数据受到过两种得州神枪手谬误的拷问。此外，它们还无法区分因果关系和随机事件，其数据挖掘式的“知识发现”甚至会让这一问题难上加难。

所有医疗专业人士都学过的准则是：首先不能造成伤害。有经验的医生对医学研究总会保持良性怀疑态度，对不喝咖啡、依赖远程祈祷和撤掉电力线都抱着“等等看”的态度。他们了解发表论文的压力和递减效应，对黑匣子数据挖掘心存质疑。我的私人医生对“依赖黑匣子算法开处方或提供医疗方案”的观点嗤之以鼻。

医疗软件程序可以辅助医生，但无法取代医生。



第11章
完胜股市（下）

> 如今，技术分析师都被称为金融工程师。我们不仅过度欣赏计算机的能力，也过于钦佩使用计算机而不使用笔和图表的金融工程师。

金融工程师不思考他们发现的模型是否合理。他们的准则是：“给我看数据就行。”其实，虽然很多金融工程师是物理学或数学博士，但其对经济学和金融学的了解过于肤浅。不过，这并没有对他们造成困扰，要说有什么影响的话，那就是无知的他们更有勇气从最不可能的地方寻找模型。

从使用铅笔的技术分析师转到使用计算机的金融工程师，对此符合逻辑的结论是要将人类彻底排除在外，数据分析的工作交给计算机做就行了。

2011年，精彩的科技杂志《连线》发表了一篇文章，全文充斥着对计算机化股票交易系统的敬畏和钦佩之情。这些黑匣子式系统被称为“算法交易者”（algorithmic traders）——由计算机根据算法来决定股票买卖，而不是人的判断。人类编写算法指导计算机，但在这之后，全靠计算机自己运行了。

有些人被唬住了。2016年，佩珀代因大学将其投资组合的10%投给了金融工程师基金，其投资总监表示：“寻找具有良好前景的公司合情合理，因为我们在日常生活中都会寻找被低估的事物，但是金融工程师策略与我们的生活毫不相干。”他认为，没有从生活中获得的智慧和常识，是支持使用计算机的论据所在。和他观点一致的大有人在。如今，美国股票交易的近1/3是依靠黑匣子式的投资算法完成的。

这些系统有的追踪股价走势，有的观察经济数据和非经济数据、剖析新闻线索。它们全都在寻找模型。一个动量算法或许会注意到，当某只股票的交易价格连续五天较高时，其第六天的股价通常也会更高；一个均值回归算法或许会注意到，当某只股票的交易价格连续八天较高时，则其第九天的交易价格通常会下降；一个配对交易算法或许会

注意到，两只股票通常会同涨同跌，当其中一只上涨而另一只没上涨时就是在提示机会来了。其他算法还使用了多元回归模型。在每一种情况下，算法都是基于数据挖掘运行的，其格言是：如果它行得通，那就好好利用。

我自己会投资，也在教授投资学，因此我决定自己尝试一下数据挖掘，看看能否找出预测股价的可靠指标。运气好的话，我的数据挖掘或许能收获“知识发现”，我可以靠此赚上一笔。

股市与天气

据报道，纽约市的天气会影响美国股市，虽然其影响随着时间的推移已经减弱，因为全美乃至全世界的股票交易已经从大厅交易演变为电子下单。

海蒂·阿蒂格搜集了25座城市每日的最高气温和最低气温数据，这鼓动我想看看能否找到某些气温，用来解释标准普尔500指数每日股价的波动情况。

我最初以为每日气温在预测股价上有局限性，因为气温随季节变化而股价不是。此外，股价具有明显的上涨趋势但气温没有（至少短短几年内不会）。尽管如此，没费多少工夫，我还是用数据挖掘找到了五个气温，很好地预测了2015年的股价。

那25座城市的最高气温和最低气温为我提供了50个可能的解释变量，以它们为基础，我可以获得：50个含有一个解释变量的模型；1 225个含有两个解释变量的模型；19 600个含有三个解释变量的模型；230 300个含有四个解释变量的模型；2 118 760个含有五个解释变量的模型。试图做一名非常投入的数据挖掘者，我将所有解释变量都推算出来了，总共是2 369 935个模型。

得到的很多模型都不错，但最好的如下：

$$Y=2361.65-3.00C+2.08M-1.85A+1.98L-3.06R$$

其中：

C = 澳大利亚科廷市，最高气温

M = 华盛顿州奥玛克市，最低气温

A = 蒙大拿州羚羊谷，最高气温

L = 蒙大拿州林肯市，最低气温

R = 怀俄明州石泉镇，最低气温

巧合的是，在第4章讨论过的科廷市和奥玛克市再次出现，不过这次最高气温和最低气温互换了。

如图11.1所示，尽管2015年下半年股市下跌，含有5个温度的模型与股价波动的吻合度很高。该模型的准确率为60%，对于预测变幻莫测的对象（如股价）来说，已经算是相当可以了。

这对“知识发现”来说又算是怎么一回事呢？有谁知道这5个小城镇的每日最高和最低气温能有助于预测股价呢？

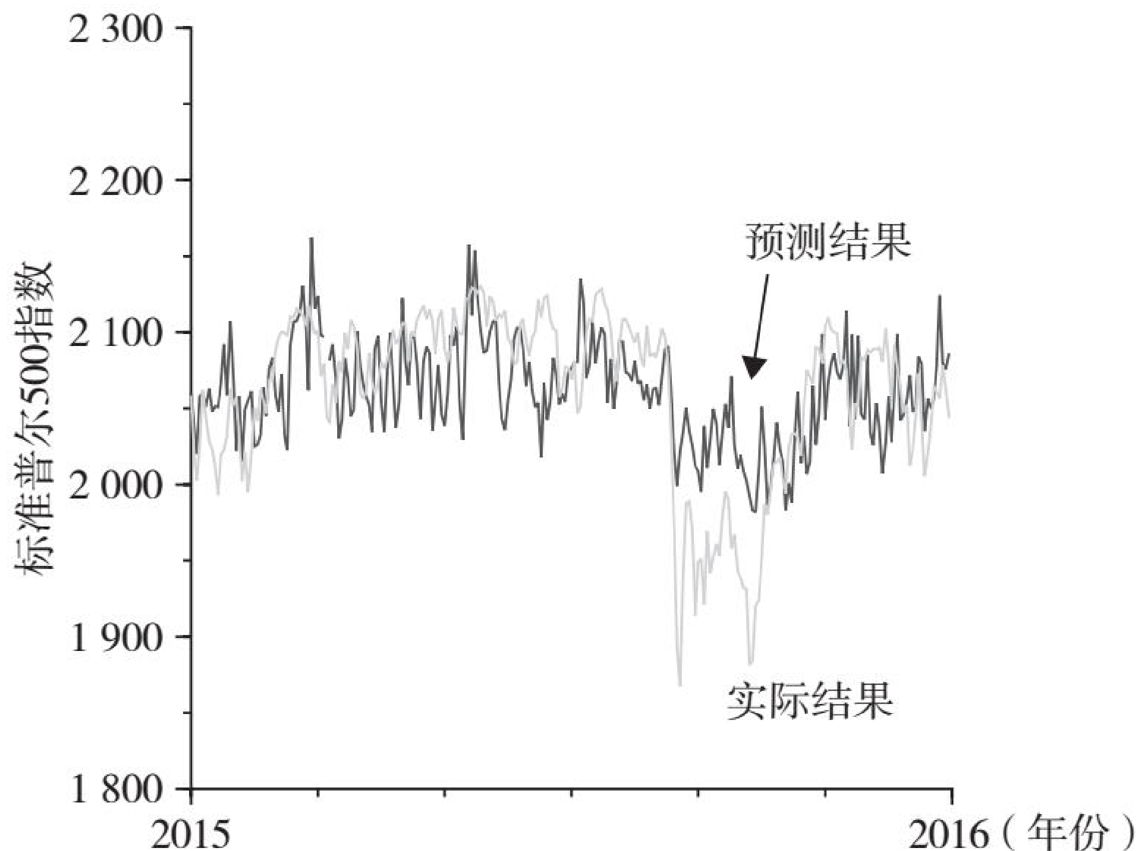


图11.1 与股价相关的“知识发现”

答案当然是它们对预测股价没有帮助。找不到合理的理由能说明标准普尔500指数与这5个城镇的最高气温和最低气温存在正相关或负相关关系，其中还有一个城市远在澳大利亚。我们能生编硬造出不切实际的说法，解释为什么每日股价取决于这些城市的消费状况，而消费状况又如何取决于这些城市的天气，但这也不过是信口雌黄而已。

先用2015年的数据推算出200多万个方程式，再从中挑出准确率最高的那一个，这就是上述模型的选择过程。由于模型建立在数据而非逻辑之上，因此我们不要指望它能较好地预测2016年的股价。如图11.2所示，2016年的预测准确率为-23%。没错，结果是个负值。当该模型预测股价将上涨或下跌时，很可能出现相反的情况。

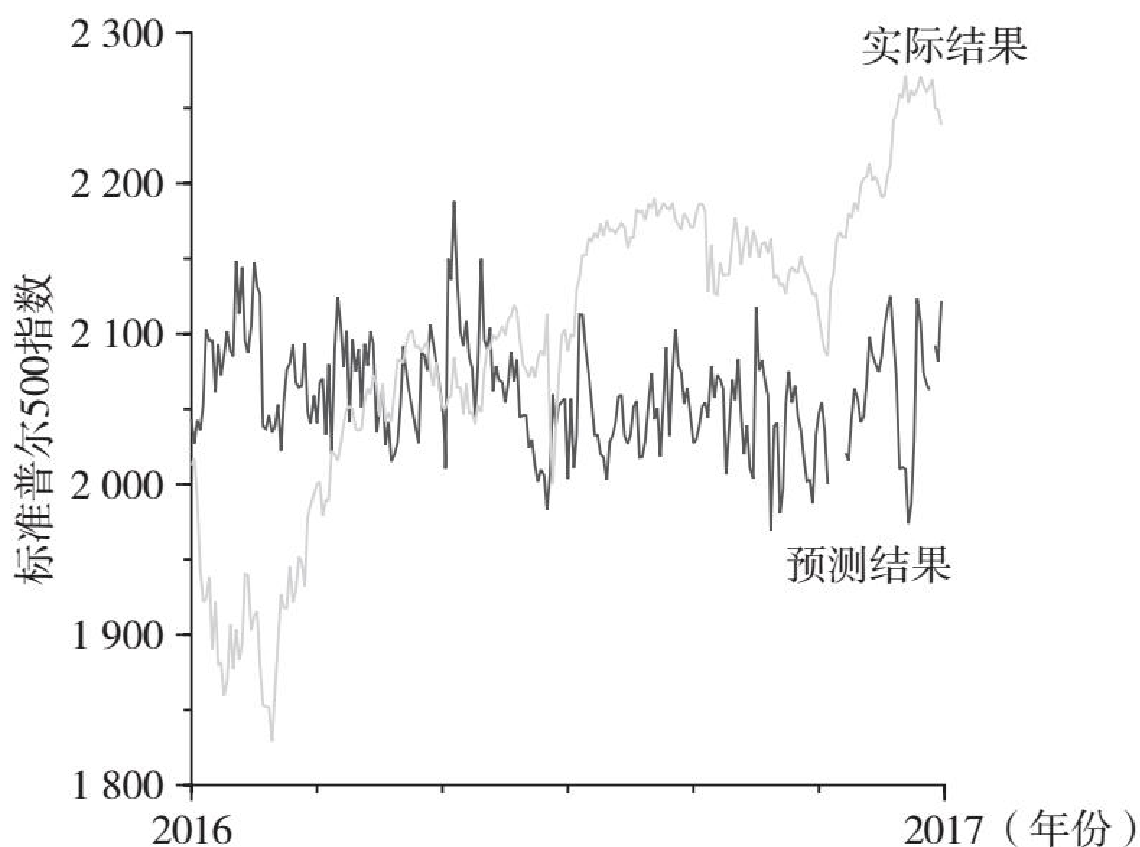


图11.2 2016—2017年的预测结果和实际结果

不断尝试

我对自己的气温模型感到失望，于是考虑用100个新的变量，尝试从1~5个解释变量的所有可能组合。现在，模型数量已接近8 000万，但是对我的数据挖掘软件来说，这个数目还是小到它能尝试每一种可能性，而无须求助主成分分析、因子分析、逐步回归法或其他有缺陷的数据规约步骤。

推算这些模型花费了数小时，所以我就止步于5个解释变量了。如果我继续推算，利用10个解释变量会得到超过17万亿个可能组合，那样的话，计算机就得花费好几天来跑数据。幸运的是，有几个5变量组合预测的股价与实际股价非常接近。最好的模型如图11.3所示，准确率高达88%，拟合数值与实际数值非常接近，实际上很难将其区分开来。



图11.3 我的5变量股价模型

我可能已经揭开了股票预测的未解之谜。你做好投资的准备了吗？

我是在2017年4月利用2015年的每日数据进行这次数据挖掘探秘的。对于含有5个气温变量的模型，我特意预留了2016年的每日数据，目的是验证我的“知识发现”。如图11.4所示，该模型对2015年的预测结果喜人，但对2016年的预测结果则一塌糊涂，它预测股价会暴跌但实际是暴涨。具体来说，该模型对于2015年预测的准确率为88%，而2016年的准确率为-52%。这个5变量模型的预测结果与2016年标准普尔500指数的实际表现存在强负相关关系，我的模型比毫无价值更糟。

图11.4 使用全新数据后的5变量模型

这是怎么回事？在某年预测效果很好的模型，怎么在下一年的预测结果会如此不尽如人意？这就是数据挖掘的本质。选择某个模型，只是因为它与所给的数据集吻合度高，这就造成这个模型与全新数据的吻合度达不到同样的水平。若要在处理全新数据时依然有效，就必须采用合理的模型。不过，数据挖掘软件无法判断一个模型是否合理。

我通过本质上是对随机数据（如澳大利亚科廷市的最高气温）进行的数据挖掘，想说服你相信这样一个事实：它们根本不会影响标准普尔500指数。我们通过逻辑推理进一步得出结论，图11.3和图11.4所示的模型，不是实质上为随机的，而是完全随机的模型。标准普尔500指数是真实数据，但那100个可能的解释变量是我用计算机随机数字生成器生成的。

还记得我曾让学生用抛硬币的方式得出虚假的股价数据吗？每只股票的起价均为50美元，然后将25次抛硬币的结果作为当天股价变化的依据，抛出正面则股价上涨50美分，抛出反面则股价下跌50美分。我在课堂上做这样的抛硬币实验是想要学生亲眼看看，明显为随机产生的数据是如何产生了看似非随机的模式的。

这次的做法也一样，不过换成了使用计算机的随机数字生成器。我将每个变量的初始值设为50，然后让电脑抛硬币来决定变量每天的变化值。若电脑抛硬币的结果为正面，则数值上升0.50；若为反面则数值下降0.50。我用计算机为每个变量的每日变化抛了25次硬币，以便得到100个虚构变量在这两年内的每日数值，将前半部分的随机数据标记为2015年，后半部分标记为2016年。

即使100个变量都是由随机游走过程产生的，在事实发生后，还是会存在有些变量的确与标准普尔500指数存在偶然的相关系数。在五变量的所有可能性中，随机变量4、34、44、64和90的组合与2015年标准普尔500指数的相关度最高。但到了2016年，该模型就完全行不通了，因为这些都是实实在在的随机变量。

黑匣子式数据挖掘无法预测这种巨大的落差，因为它不能评估自己发现的模型是否具有逻辑基础。

预留方案

现在，可能有人会说，既然从该模型对2016年预测结果的糟糕程度就可以看出标准普尔500指数和我的随机变量之间不存在任何真正的关系，那么我们就可以利用样本外测试来区别偶然的相关系数和真正的因果关系。挖掘部分数据，寻找“知识发现”，然后通过有目的的暂时预留的数据来测试所发现的模型以验证结果。原始数据有时被称作“训练数据”，预留数据被称作“检验数据”或“验证数据”。另一种叫法为样本内数据（用以发现模型的数据）和样本外数据（用以验证模型的全新数据）。在利用气温和随机变量预测标准普尔500指数的例子中，模型是用2015年的数据推算得出，用2016年的数据进行验证的。预留出2016年的数据，正是为了这一目的。

不断询问模型是否运用全新数据验证过是一个很好的想法。大肆搜集数据以发现模型，再用相同的数据来验证模型的做法绝对没有说服力，这些数据都是为了找到模型而被掠夺来的。因此，预留验证数据来检验无中生有、生编硬造来的模型肯定不失为好方法。

然而，不知疲倦的数据挖掘可以确保某些模型与训练数据和检验数据的吻合度都很高，即便该模型根本不合理。正如有的模型肯定与原始数据吻合，有的仅仅是运气好，也能与预留数据吻合。发现同时符合原始数据和预留数据的模型，只不过是另一种数据挖掘形式。我们要找的不是符合半数数据的模型，而是符合所有数据的模型。为了符合数据而挑选的模型，无论是符合半数还是所有数据，都不能指望它与其他数据的吻合度一样高。这么做解决不了问题。

为了说明这一点，接下来看看我为了解释标准普尔500指数的波动而创造出的100个随机变量。共有100个含有一个变量的模型：随机变量1、随机变量2……对于每一个变量，我都利用2015年的每日数据，来推算出吻合度最高的模型。以随机变量1为例：

$$Y = 2113.62 - 0.5489R1$$

该模型的准确率（标准普尔500指数的预测值和实际值之间的相关系数）为28%。但我用此模型预测2016年的标准普尔500指数时，其准确率竟为-89%。该模型预测标准普尔指数会上涨，但实际上该指数下跌了，反之亦然。

我把100个可能的解释变量统统用上，反复尝试，让模型与2015年的数据吻合，再用2016年的数据验证，结果如图11.5所示。对于2015年的

数据，由于它们被用以推算模型，所以准确率不可能小于0，因为该模型总能完全忽略解释变量，从而得到准确率为0。结果显示，使用样本内数据且含有1个变量的模型的平均准确率为27%。

对于预留下来用以验证模型的2016年的数据，其准确率为正值和负值的可能性相等，因为毕竟它们是与股价毫无关系的随机变量。我们预计，股价和任何随机变量之间的平均相关系数均约为0。对上述特定数据来说，样本外数据的平均准确率碰巧为-4%。

尽管如此，样本外数据的准确率还是会碰巧与某些模型存在强正相关系数，与其他模型存在强负相关系数。如图11.5右上角所示，有几个模型的2015年样本内数据和2016年样本外数据的准确率都很高。具体来说，有11个使用2015年拟合数据的模型相关系数高于0.5，其中5个在使用预留数据时的相关系数高于0.5。这五个模型都通过了样本外数据验证测试，尽管它们只是与股价完全没有关联的随机变量。

若使用更多解释变量，准确率还会上升。我又重复了一次实验，推算了4 959个可能的双变量模型。随机变量57和59的吻合度最高：

$$Y = 2100.46 + 3.4612R57 - 4.8283R90$$

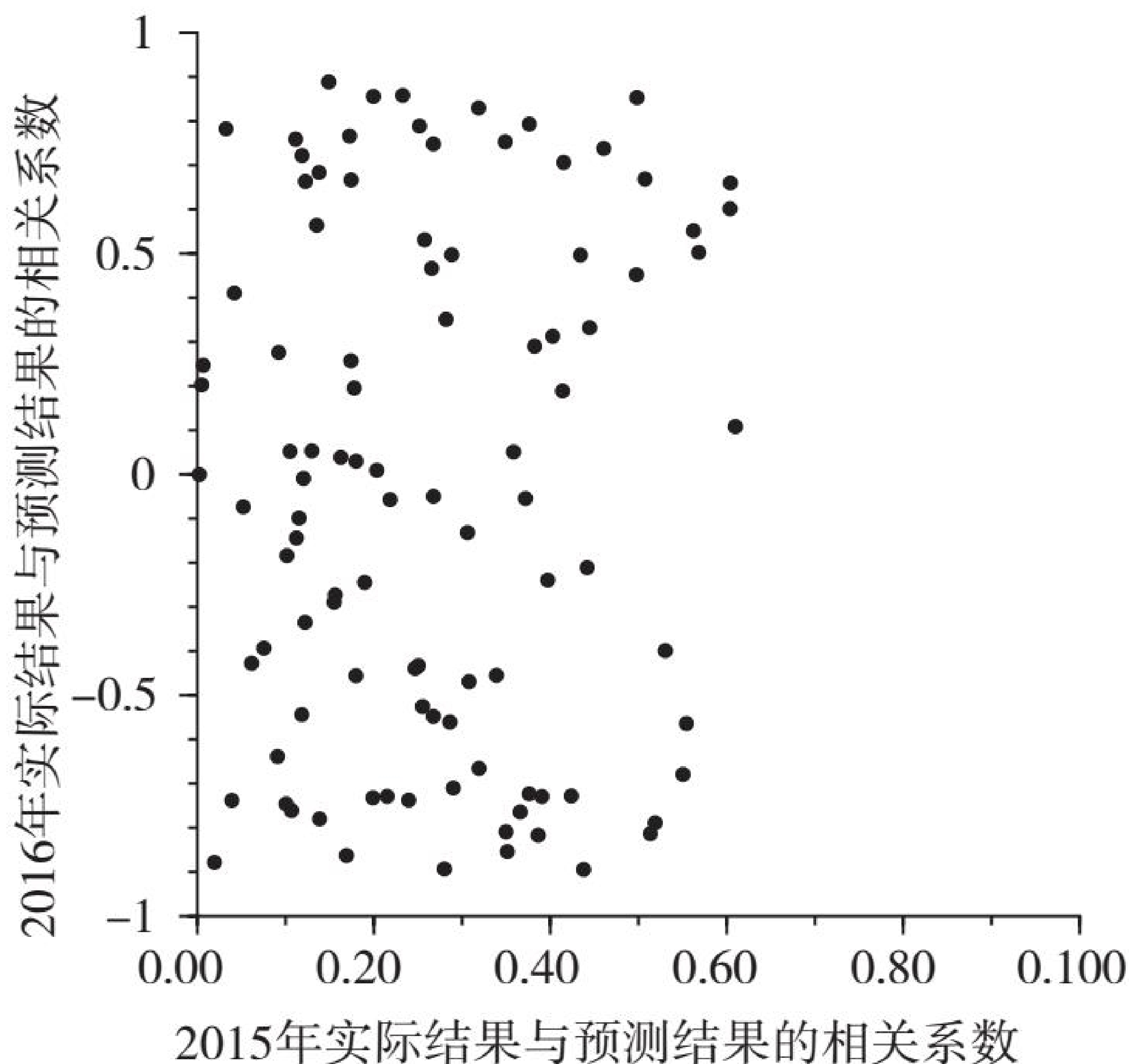


图11.5 100个单一变量模型的样本内和样本外数据吻合情况

这个模型的样本内数据准确率竟高达79%，但是样本外数据准确率为-56%。这个双变量模型与2015年的数据高度吻合，但是对2016年的预测结果却与实际值呈负相关关系。尽管有这个缺点，但更多的数据挖掘肯定会找到既符合2015年的训练数据，又符合2016年的验证数据的模型。

使用双变量的模型，2015年回测数据的平均准确率为40%，而2016年预留数据的平均准确率为-1%。图11.6为2015年和2016年准确率之间的关系。

这4 950个模型把图表变成了巨大的斑点。有很多模型（如随机变量57和90）与2015年数据吻合度高，但与2016年数据的吻合情况一塌糊涂。同时，也有很多模型与这两年的数据都非常吻合，有时，与2016年数据的吻合度甚至高于2015年。这就是偶然的本质，这些都是偶然得出的变量。

有46个模型的2015年准确率为70%，其中11个模型的2016年准确率为70%。这11个模型都通过了验证测试，但它们对预测其他年份的股价还是没有效果，如2017年。

图11.6 4950个双变量模型的样本内和样本外数据吻合情况

有一个使用随机变量14和74的模型，其2015年准确率为70%，而对于2016年验证数据的准确率竟然达到88%！如果我们对此不够了解，可能还以为自己取得了什么重大发现。然而事实是，人们总能找到同时符合样本内和样本外数据的模型，即使这些数据都不过是随机噪声。

对含有更多解释变量的模型来说，情况则有过之而无不及。若变量增加，可能的模型数量会呈现爆炸式增长，找到符合训练数据和预留数据的模型的确定性也会更大。含三个变量的可能模型有161 700个，含四个变量的可能模型有3 921 225个，含5个变量的可能模型有75 287 520个。

随着可能性越来越多，图表会密密麻麻地布满圆点（如图11.6所示）。但是，原则仍然成立。从中肯定能找到很多模型同时与2015年和2016年的数据吻合。

例如，最佳的五变量模型的2015年样本内数据准确率为88%，2016年样本外数据准确率为-52%。然而，有些5变量的模型碰巧与2015年的吻合度高，有些是与2016年的吻合度高，还有些在这两年的吻合度都很高。我的数据挖掘软件识别了11 201个5变量模型，这些模型与2015年标准普尔500指数的实际值和预测值之间的相关系数至少为85%，其中有109个模型的2016年准确率高于85%，49个模型的2016年准确率高于90%。如果我再尝试更多变量，我的数据挖掘软件肯定会发现对两年的准确率都高于90%，甚至高于95%的模型。

这不是“知识发现”，而是偶然发现。

如果我们搜遍股价数据就是为了找到不合理的系统以完胜股市，几乎可以肯定的是，我们会因此更穷。

真正的数据挖掘

Quantopian（众包型量化投资平台）网站为想要成为投资大神的人提供编写其交易算法的空间，再用历史数据回测，看看这些算法会带来多大回报。听起来很合理。不过，我们知道，数据挖掘总能找到在挖掘期内获利的算法。我们还知道，没有逻辑基础的算法在使用全新数据时的表现通常会让人大失所望，无论它们的回测结果有多好。

Quantopian平台有意思的一点在于，尽管这些算法的细节没有公开，但任何人都可自主采用过去任何时间段的数据进行验证。此外，每个算法都有时间标记，显示该算法的最后版本是于何时发表在Quantopian平台上的。

有外部团队检验了该平台将近1 000个股票交易算法，这些算法均发表于2015年1月1日到6月30日。每个算法都利用2010年至发表前的数据进行回测（训练期），然后再用发表后到2015年12月31日的全新数据进行检验（验证期）。结果发现，训练期和验证期的收益之间存在很小但是统计学意义显著的负相关关系。大写的尴尬！

趋同交易

卖空股票是指卖出从其他投资者手中借来的股票。有时候还必须回购股票（希望是以更低的股价）还给投资者。现在，假设你能以90美元的价格买入一只股票，并以100美元的价格卖空同一只股票。如果这两个股价趋同于110美元，那么你的第一只股票的收益为20美元（90美元买入，后以110美元卖出），第二只股票损失了10美元（100美元卖出，后以110美元回购）。因此，你的净收益为10美元，即两个初始股价之差。

相反，如果这两个股价趋同于80美元，那么你的第一只股票损失了10美元（90美元买入，后以80美元卖出），第二只股票收益为20美元

（100美元卖出，后以80美元回购）。因此，你的净收益为10美元。

这就是所谓的“趋同交易”，因为你赌的不是两只股票的涨跌，而是股价会趋向一个共同的价格。

荷兰皇家壳牌集团

1907年，荷兰皇家石油公司（总部位于荷兰）和英国壳牌运输和贸易公司（总部位于英国）合并经营，联手对抗约翰·D.洛克菲勒的标准石油公司——全球最大的炼油公司。荷兰皇家石油将专注于生产，英国壳牌则专注于销售，合并经营之后，这两家公司或许还能存活下去。

让人好奇的是，根据双方协议，荷兰皇家石油和英国壳牌均保留各自目前的股东，这两只股票也继续在各家证券交易所进行交易，不过，所有收益和支出都合并到母公司荷兰皇家壳牌集团（荷兰皇家石油占股60%，英国壳牌占股40%）。集团全部收入的60%归荷兰皇家石油，40%归英国壳牌；集团派发的全部股息的60%归荷兰皇家石油的股东，40%归英国壳牌的股东；如果集团被出售，收入的60%归荷兰皇家石油的股东，40%归英国壳牌的股东。

无论英国壳牌的价值为多少，荷兰皇家石油的价值都要比它高出50%。如果股市对两者股票的估值正确，荷兰皇家石油的股票市值应该总会比英国壳牌的高出50%。但事实并非如此！

图11.7为1957年3月13日（两只股票首次在纽约证券交易所进行交易）到2005年7月19日（两家公司完全合并，股票停止单独交易），荷兰皇家石油与英国壳牌的股票市值比率。



图11.7 荷兰皇家石油与英国壳牌

荷兰皇家石油的股价几乎从未刚好比英国壳牌的高50%，有时会高40%，有时会低30%。从整个时间段来看，两者的实际市值比与正确的理论市值比（1.5）之间的百分差有46%的时间高于10%，有18%的时间高于20%。

这种情况非常适合趋同交易。当荷兰皇家石油的交易价与英国壳牌的交易价之比高于1.5时，投资者可以买入英国壳牌，卖空荷兰皇家石油，赌这个溢价会消失。

1997年，美国长期资本管理公司就这么做了，当时的溢价从8%涨到10%。该公司买入英国壳牌价值11.5亿美元的股票，卖空荷兰皇家石油价值11.5亿美元的股票，坐等市场修正股价。该公司拥有全明星阵容的管理团队，包括两名荣获1997年诺贝尔奖的金融学教授，这是很聪明的一招，其依据是具有说服力的逻辑，而不仅是偶然发现且毫无意

义的统计学模式。市值比率最终应该达到1.5，长期资本管理公司会从这次机智的对冲交易中获利。

然而，正如凯恩斯在大萧条期间观察所得：

> 这套长期理论在误导当前事物。从长期来看，我们都难逃一死。经济学家为自己设置了过于容易、过于无用的任务，如果遇上狂风暴雨，他们唯一能告诉我们的只有：暴风雨过后，大海会恢复平静。

凯恩斯嘲讽的观点是：从长远来看，经济发展会趋于平静，想找工作的人总会找到工作的。他认为，短期的经济衰退风暴比假想的长期平静更加重要，或许没人能够看到那个长期的到来。股市也是如此。从长期来看可获利的趋同交易，从短期来看却会带来灾难性后果。

1998年初，长期资本管理公司的净价值接近50亿美元。同年8月，一场始料未及的风暴来袭。俄罗斯未能偿还债务，并且察觉到整个金融市场的度量风险都在提高。长期资本管理公司在很多不同市场都下了赌注，猜测大部分的风险溢价将下降。但自俄罗斯未能偿还债务后，到处都出现了风险溢价上涨，长期资本管理公司遇到了麻烦，而且是很大的麻烦。

该公司争论道，一切都是时间的问题，等时候到了，金融市场就会恢复到正常水平——暴风雨终将过去，大海会再次平静——但是，该公司已经等不起了。它下的大赌注和借款之间的杠杆过高——若能偿清就尚好，否则会导致灾难性的后果。8月21日，该公司损失了5.5亿美元，整个月下来共损失了21亿美元，将近其净价值的一半。

长期资本管理公司努力筹集更多资金，期待熬过这次风暴，但贷方已成惊弓之鸟，不愿再给该公司放贷，还想着讨回先前的借款。

凯恩斯不仅是大师级经济学家，还是传奇般的投资家。他曾告诫：“市场保持非理性状态的时间，可比你保持有偿还能力的时间更长。”可能市场对俄罗斯未偿还债务的反应过度了，也可能长期资本管理公司最终会转亏为盈。但是，它保持有偿还能力的时间，不足以让它见证这一刻的到来。

该公司不得不对其持有的荷兰皇家壳牌集团股票进行平仓处理，当时，荷兰皇家石油的溢价不降反升，超过了20%。长期资本管理公司在

这笔交易中损失了1.5亿美元。

同年9月23日，沃伦·巴菲特给该公司传真了一封信件，提出要以2.5亿美元收购该公司，约为其年初净价值的5%。这次出价是“要卖就卖，不卖拉倒”型，截止时间为当天中午12点30分，也就是传真后的一个小时。该公司最后没有接受出价，开始为自己准备“后事”。

纽约联邦储备银行担心长期资本管理公司未偿还债务会引起多米诺效应，触发全球金融危机。于是，纽约联邦储备银行携手长期资本管理公司的债权人接管该公司并清算其资产。债权人收回了贷款，公司创始合伙人损失了10.9亿美元，其他投资者则花大价钱上了一课，了解到了杠杆的力量。

注意看图11.7，2005年，溢价最终的确消失了，当时荷兰皇家石油与英国壳牌合并，荷兰皇家石油的股东拿到了合并公司60%的股份，英国壳牌的股东则拿到了其余的40%。荷兰皇家壳牌集团这次的交易确实是明智之举，合情合理且最后也成功获利。不幸的是，长期资本管理公司的那些交易就欠缺考虑，最后迫使自己不得不过早清算了荷兰皇家壳牌集团的股票。

股市价格有时稀奇古怪，荷兰皇家壳牌集团的错误定价就是非常有说服力的例证。无论英国壳牌股票的“正确”价值是多少，荷兰皇家石油总会多出50%，然而股市价格时高时低，为有利可图的趋同交易创造了机会。然而，这个例子还说明，即使是由行业顶级人士正确无误地完成的趋同交易，也是有风险的，因为趋同所需时间可能比预期更长。而没有逻辑基础的趋同交易就更加危机四伏了。

黄金白银比率

20世纪80年代，大名鼎鼎的投资顾问公司Hume & Associates（休姆联合公司）制作出《超级投资者档案》（The Superinvestor Files），向全美宣传，普通投资者靠它就能获得非常可观的利润。订阅用户每月会收到一份印刷精良的50页册子，每本25美元，外加合计2.5美元的邮费和处理费。

回想起来，本应显而易见的是，如果这些策略像广告宣传的那样有赚头，该公司利用这些策略可以比推销册子挣更多的钱。然而，容易受

骗、贪婪的投资者忽视了这一点，反而希望花上25美元和合计2.5美元的邮费和处理费，就买到成为百万富翁的秘诀。

其中一个超级投资者策略是基于黄金白银比率（gold/silver ratio,GSR），即每盎司黄金和白银的价格比率。1985年，黄金的均价为317.26美元，白银的均价为5.88美元，则GSR为317.26美元/5.88美元 = 54，也就是说，每盎司黄金的价格是白银的54倍。

1986年，休姆写道：

> GSR在过去七八年内的波动幅度较大，1980年低至19∶1，1982年高达52∶1，到了1985年又升至55∶1。但是，你也能清晰地看到，它总是（注意，总是）会回到34∶1~38∶1的范围。

图11.8证明了1970—1985年的GSR在34~38的范围内波动。



图11.8 1970—1985年的GSR

GSR策略是在GSR处于异常高的状态时，卖金买银；处于异常低的状态时，则买金卖银。采用期货合约使这些交易产生了巨大杠杆，有可能

获得暴利。这是一次趋同交易，因为投资者赌的不是金银价格的涨跌，而是两者比率会趋同于其历史比率。

一盎司黄金的价格应是白银的36倍，其中原因毫无规律可循。黄金和白银不像鸡蛋，可以买一打或半打，如果价格有偏差，消费者会买更便宜的鸡蛋；也不像玉米、大豆，如果玉米相对大豆的价格上涨，农民就种更多的玉米。

最终显示，1983年GSR上涨到38后，直至2011年，也就是28年后，才回落。如图11.9所示，GSR保持在34~38的范围内只是短暂出现的巧合，不是这个超级策略的基础。期货合约可能成倍扩大损失和收益，而在1985年下注GSR则会产生灾难性的后果。



图11.9 1970—2017年的GSR

其他趋同交易

早期趋同交易只依据简单的模式进行，如GSR，在价格图上一目了然。现代计算机能搜遍大量数据库，寻找更不易察觉和复杂的趋同交易。

如果两个价格之间的相关系数为0.9，价格开始往不同方向移动，交易算法可能就会判断这一历史关系会重现。

即使计算机发现了模型，GSR交易中也存在同样的问题，即无理论支撑的数据存在隐患。趋同交易需要合理，因为如果找不到所发现模型的根本原因，该模型出现的偏离也就没有理由自我修正。统计学相关系数可能是偶然出现的模型，转瞬即逝。

荷兰皇家壳牌集团的趋同交易就有很合理的基础，但是长期资本投资公司破产的原因是，将大量赌资压在了根本原因不具有说服力的相关系数上。例如，法国和德国各种利率之间的关系与风险溢价。一名经理后来痛惜道："我们公司的学术大师在加盟时毫无交易经验，就这样开始建模。鉴于自己做出的假设，他们的交易看似不错，但常常连简单的可信度检验都无法通过。"

讽刺的是，观察荷兰皇家石油和英国壳牌在任何一两年内的每日股价，黑匣子式的交易算法都不会识别出它们股价的比率应该是1.5。它会漏掉其中一次合理的趋同交易。

图11.10所示的是一个更近时期的趋同交易机会。在2015年和2016年大部分时间，这两只股票的股价比率波动范围的平均值为0.76。虽然价格比率相对于0.76时高时低，但总是会回到明显的均衡值。

2016年8月25日，该价格比率突破1，表明这是卖出一股买入另一股的好时机。可惜的是，如图11.11所示，该比率并没有回到自然均衡值0.76，而是继续上升，翻了一倍多。2016年11月3日，其峰值高达2.14，而后才稍有回落。

或许，该比率终有一天会回到0.76，也可能不会。

我如何得知？因为这些是我用随机数字生成器捏造出的数据。我再一次老调重弹，使用了计算机随机数字生成器。将起始股价设为50美元，然后用电脑抛25次硬币，正反面分别加减50美分，得到10只虚构股票两年内的每日股价数据。

随后，我又观察了配对股价的比率，也没花太长时间。图11.10和图11.11的数据，就是随机股价2和随机股价1的比率。

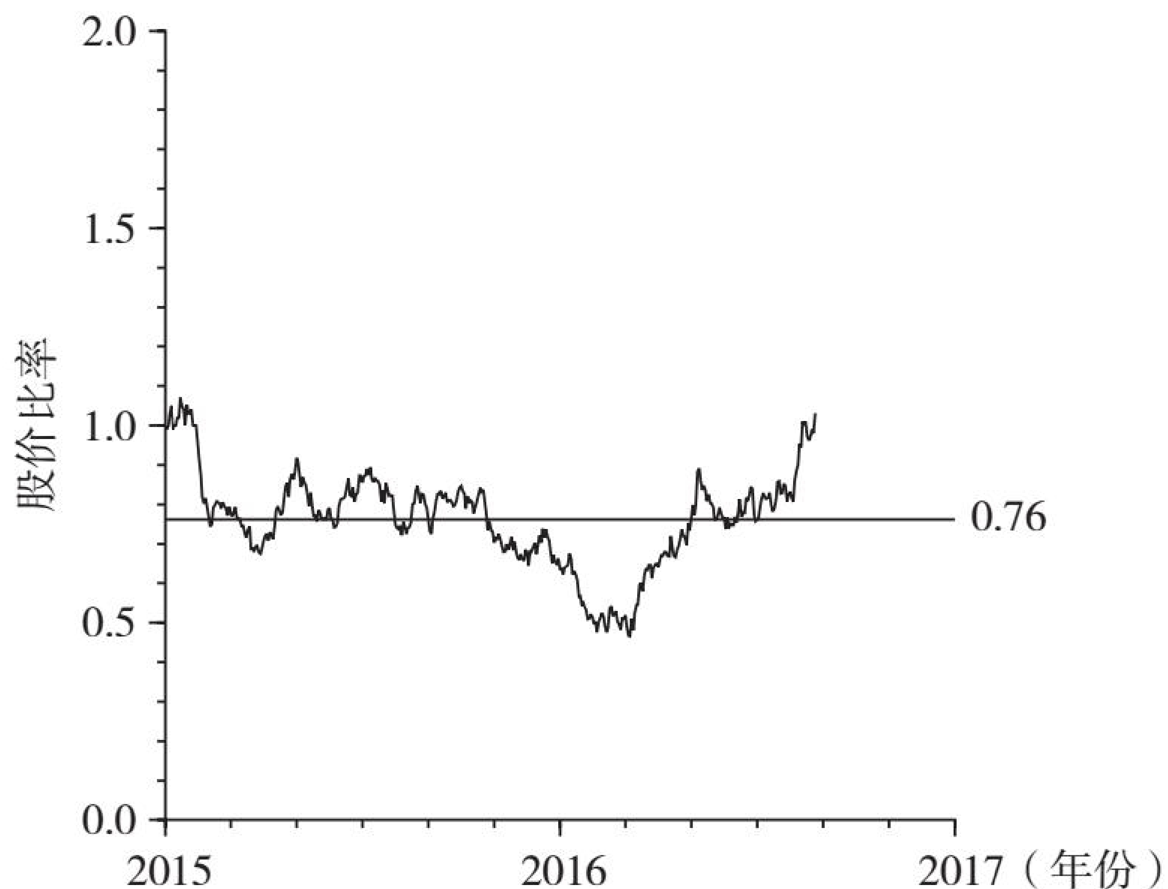


图11.10 趋同交易

图11.11 尴尬的局面

比率初始值为1，是因为每个虚假股价的起价均为50美元。抛硬币碰巧得出的比率围绕0.76来回移动了一年半。之后，比率突然暴涨，接着有所回落。该比率下一步会如何变动？我不知道，这完全取决于计算机的随机数字生成器。

随机股价1和随机股价2完全独立。它们的共同点仅在于起价均为50美元。在这之后，股价1每日的变化都由电脑的25次抛硬币决定，股价2也是如此。每个价格都跟随随机游走程序，涨跌可能性一样大，完全独立于其他价格路径。但是，它们的价格比率似乎都停留在0.76附

近，不会长时间偏离，很快又会再次恢复。后来，随机游走程序突然将比率带得远离0.76，可能再也不会恢复。

要注意的是，即使数据完全是随机生成的，还是会出现适合进行趋同交易的情况。但这并不代表每次潜在的趋同交易都是随机噪声。要提醒大家的是，我们无法像黑匣子式的数据挖掘软件那样，仅靠观察数据，就能分辨出趋同模型反映的情况是真实的还是偶然的。计算机对于判断趋同模式是否具有逻辑基础完全无能为力。只有人类才能判断关系的形成理由是否具有说服力。对荷兰皇家壳牌集团来说，这个答案是肯定的。但对GSR来说，就不是了。

高频交易

有些算法被用以进行高频交易，使其买卖速度快到超乎人的想象。计算机可能会注意到，只要股价下跌的股票数量在接下来的140秒内比股价上涨的股票数量超出8%，标准普尔500指数的期货价格通常也会上涨。计算机将这一指标存档待用。当同一信号再次出现时，计算机便发动猛攻，立刻买入数千手标准普尔500指数期货，随后又迅速卖出。《连线》杂志对这些自动化系统赞不绝口，认为它们“比所有人类都更高效、快速和聪明”。更快速，确实是的；但更聪明，并非如此。

投资公司花费几十亿美元建立靠近股市的交易中心，使用光纤网线、微波塔台和激光通信线路，将芝加哥、纽约、伦敦、法兰克福和东京的信息传播与交易下单时间缩短至毫秒和纳秒。例如，纽约证券交易所和芝加哥商品交易所之间的一连串微波塔台，能在9毫秒内往返发送距离超过700英里的买卖订单。为什么要这么做呢？

第一个目的是，利用可察觉到的定价差异。假设IBM的股票在一家交易所的每股买入价为200.0000美元，在另一家交易所的每股卖出价为200.0001美元。发现这一异常现象的计算机程序会以200.0000美元尽量多地买入，为的就是在一毫秒后再以200.0001美元卖出，直到这一价格差异消失。每股0.0001美元的收益并不多，但是如果在一秒内完成数百次或数千次交易，就能产生非常可观的年收益。

在理性的世界里，资源不会浪费在这些无意义的事情上。不同交易所的股价出现如此细微的差异，这真的重要吗？差异定价持续了9毫秒，而不是10毫秒，这真的要紧吗？

极速交易的第二个目的是，比普通投资者更快一步下单。如果杰里下单，以当前市场价格买入1 000股股票，极速运转的交易程序可能会先买入，又在毫秒之间卖给杰里，一来一回每股赚取1美分的利润。以每股1美分的收益交易1 000股，就获得了10美元的收益。如此不断重复，利润可以达到数百万美元。计算机坑骗杰里，让他每股多付1美分，这给社会带来了什么经济利益呢？毫无利益可言，只不过是个计算机化的扒手偷了钱，受害者甚至还蒙在鼓里。

更根本的是，用超级智能程序来运行极速交易程序，而不是将其用在大有裨益的地方，会带来什么经济利益？建立交易中心，布好传输线路来加快股市下单，而不是将这些资源用在大有裨益的地方，会带来什么经济利益？

极速交易反而会雪上加霜，导致经济损害。

如果有人让计算机寻找有可能获利的模型（无论所发现的模型是否合理），然后在模型重现时买入或卖出，计算机会唯命是从（无论这个模型是否合理）。的确，计算机背后会有人吹嘘，他们真的不知道为什么自己的计算机会自行决定交易。毕竟，计算机比他们更聪明，不是吗？他们该做的不是自吹自擂，而是自求多福。

指令克隆问题也提高了黑匣子式高频投资的风险。如果软件工程师给几百台计算机下达相似的指令，就会有数百台计算机在同一时间竞相买卖同一只股票，广泛影响金融市场的稳定。《连线》杂志值得赞扬的是，它认识到了无人监管的计算机一致运作存在危险：“最糟糕的情况是，无人监管的计算机变成难以捉摸的反馈循环……最终击垮了计算机系统。”

闪电崩盘

2010年5月6日，美国股市受到著名的“闪电崩盘”的冲击。当天，投资者都担心希腊债务危机，一名焦虑的互惠基金经理设法卖出41亿美元的期货合约，以对冲其投资组合。他的思路是：如果市场下跌，基金股票投资组合的损失可以用期货合约的收益抵消。这一看似谨慎的交易，不知怎么就触发了计算机。计算机买入大量该基金卖出的期货合约，然后又迅速卖出，因为它们不喜欢长期持有头寸。期货价格开始下跌，于是计算机决定加大买入卖出的数量。受到刺激的计算机疯

狂进行交易，自买自卖基金的期货合约，就像一个被丢来丢去的烫手山芋。

没有人确切知道计算机为什么会突然一发不可收拾。记住，就连计算机背后的人也不明白，计算机为什么会进行交易。在15秒的间隔时间内，计算机跟自己完成了2.7万次期货合约交易，占总交易量的一半，在疯狂的15秒结束后，净购买量只有200份合约。这一疯狂交易扩散到了常规股市的交易大厅里，卖出的订单淹没了潜在的买家。道琼斯工业平均指数在5分钟内下跌近600点。坚如磐石的蓝筹股宝洁的股价也在不到4分钟内下滑了37%。有些计算机为苹果公司、惠普公司和知名拍卖行苏富比支付的每股股价超过10万美元。还有些计算机将埃森哲咨询公司的股票和其他主要股票以每股不足1美分的价格卖出。这些电脑都没有常识，它们完全不知道苹果公司和埃森哲咨询公司的价值。只要算法下达指令，它们就盲目地买卖。

直到期货市场的内置安全卫士中止所有交易5秒，这一疯狂局面才得以落幕。令人难以置信的是，这短短的5秒，就足以说服计算机停止它们的疯狂交易。15分钟后，市场恢复正常，道琼斯工业平均指数短暂暴跌600点也只是梦魇般的回忆。

在那以后还发生过“闪电崩盘”，未来可能会出现更多。令人匪夷所思的是，2013年8月30日，宝洁再次于纽约证券交易所遭遇一次微型“闪电崩盘”，之所以这么说，是因为并没有对该交易所其他股票产生特别大的影响，宝洁在其他交易所的股票也没有受到特别大的影响。

莫名其妙的是，纽约证券交易所的约200次交易，包括涉及宝洁股票在1秒内完成的约25万股交易，触发股价下跌了5%，从77.50美元降至73.61美元，随后在不到1分钟内即恢复。有个运气好的人，恰巧在正确的时间，出现在正确的地方，买入了6.5万股该股票，立刻赚了15.5万美元。为什么会发生这种情况呢？无人知晓。

虽然交易所启动安全卫士限制了进一步的“闪电崩盘”，但这也说明了黑匣子式投资算法的根本问题。这些计算机程序不知道每只股票（或任何其他投资）是真的廉价还是昂贵，甚至没有试图估算这只股票的真正价值。这就是为什么计算机程序可能以10万美元每股的价格买入苹果的股票，而仅以1美分每股的价格卖出埃森哲的股票。

底线

计算机没有常识或智慧。它们能识别统计学模式，但无法判断所发现的模式是否有逻辑基础。20世纪80年代，当黄金和白银价格的统计学相关系数被发现时，计算机程序怎么可能察觉得到这个统计学相关系数是否具备合理的理论基础呢？在宝洁股票价格瞬间下跌5%时，计算机又怎么能判断这次暴跌是有理有据，还是荒谬离谱的呢？

“人非圣贤，孰能无过。”但是人也有潜力识别那些错误，避免被计算机模型所诱惑。

我曾经的一名学生创立了一家成功的基金公司，采用的是投资其他投资基金的策略，而不是直接买入股票、债券和其他资产。他勤勉努力，采访了数千名投资者和基金经理。最后，他确定了金融工程师对冲基金的四种类型（有些基金经理会结合多种策略使用）。

1. 纯套利。利润来自交易等同或接近等同资产，通常为高频交易。例如，在两所不同的交易所的同一只股票。其利润一般很小，但稳定，风险小。

2. 市场制造者。利用股价差异，例如，在不同交易所的相似证券以极小差价进行交易。获利可观，但风险在于，交易不按照预期价格执行，尤其在交易所于不同时间和不同日期开盘时（受假期影响）。

3. 统计学套利。使用数据挖掘算法来识别有可能是有利可图的交易基础的历史模型。利润丰厚，但风险也很大。例如，在一家航空公司买入股票，到了另一家便卖出。

4. 基本面量化。采用基础数据（如股价/收入），有助于支持具有某些特征的股票，同时避开或卖空具有相反特征的公司。

他对金融工程师的总体评估是：“如今有几千名‘金融工程师’投资者和互惠基金。只有小部分能够实现极好的长期获利。正如音乐家，可以靠不止一种类型的音乐取得成功。爵士、摇滚和乡村音乐艺术家的演出门票都会销售一空，同时还有几千名其他类型的音乐家在更廉价的夜店或街道转角演奏。”



结语

> 我们生活在一个不可思议的历史时期。计算机革命比工业革命给人们的生活带来了更加翻天覆地的变化。我们可以使用计算机来实现过去无法完成的目标，计算机也为我们打开了很多崭新的大门。

我很迷计算机，你可能也有同感。但是，我们不应该让自己对计算机的喜爱，蒙蔽了对它们的局限的认知。没错，计算机储存的事实数据比我们多，记忆力比我们好，计算速度比我们快，还不会像我们那样疲倦。

机器人完成重复单调任务的能力远超人类，如拧螺栓、播种、搜索法律文件、接受银行存款和分配现金。计算机能识别物体、画画和驾车。你肯定还可以想出计算机其他让人惊叹的，甚至是超人类的壮举。

因为计算机能够极其出色地完成任务，所以很容易让人认为它们肯定是高度智能化的。然而，在完成特定任务方面大有用处与拥有通用智能是两码事。通用智能可以将从一次任务中吸取到的教训和习得的技能，运用于更加复杂或完全不同的任务。有了真正的智能，技能便可信手拈来。

计算机非常强大，而且越来越完善，但是计算机算法的设计，仍然是完成定义明确的琐事所需要的、适用范围非常狭窄的能力，而不是像通用智能那样可以通过评估事情现状、起因和后果，来处理不熟悉的情境。人类能够将通用知识运用到特定情境中，再借助特定情境来改善自己的通用知识。如今的计算机还无法做到这一点。

人工智能和人脑的真正智能完全不是一码事。计算机并不知道词语的意思，因为它无法像我们一样感知世界。它不知道真实世界是什么，缺少人类在现实生活中积累所得的常识或智慧；无法构想出有说服力的理论学说，也无法做出归纳推理或长期规划；没有情绪、感觉和灵感，这些都是创作扣人心弦的诗歌、小说或电影剧本所必不可少的。

或许有一天，计算机会拥有类似人类的真正智能，但这并不是因为计算机内存更大或处理速度更快。这不是量变的问题，而是质变产生的不同方式——找到方法让计算机获取通用智能，使其可以在不熟悉的情境中灵活运用多种方式。

我想澄清一点，这不是在批评计算机科学家。他们都才智过人，也付出了大量辛勤汗水。计算机科学家的工作难度极大，并且大有裨益。还有更多需要完成的工作，难上加难。

模仿人脑是一项艰巨的任务，不能确保一定会成功。不过，还是有一些传奇式的例外，如美国电话电报公司的贝尔实验室、洛克希德·马丁公司的“臭鼬工厂”和施乐公司的帕克研究中心，但是很少有企业愿意支持与脑力有关、短期无回报的研究。一些有用且能立即获利的项目对它们来说更具吸引力。

我不知道，开发出可与人类相媲美的通用智能的计算机需要多长时间。我猜测，至少也需要几十年。可以肯定的是，那些声称计算机已经拥有通用智能的说法都是错的。我也不相信那些人给出的特定日期，如2029年。同时，请保持对牵强附会的科学小说场景的怀疑态度，也小心提防夸大宣传人工智能产品的企业。

挖掘大数据风行一时，但数据挖掘是人为，而非智能。当统计模型分析大量可能的解释变量时，可能关系的数量就会暴增。有1 000种可能解释变量的多元回归模型，10个输入变量存在近1万亿个万亿的可能组合。若有1万个可能解释变量，则10个输入变量存在超过10亿万亿个万亿的可能组合。难以想象，若有100万个可能解释变量，会存在多少种可能组合。

如果把很多可能变量都考虑在内，即便所有都只是随机噪声，部分组合也一定与我们试图预测的对象高度相关，如癌症、信用风险和岗位适用性。偶然会出现真正的“知识发现”，但是，考虑的解释变量越多，所发现关系只是偶然出现且转瞬即逝的可能性就越大。

统计学证据不足以辨别真知灼见和虚假信息。只有逻辑、智慧和常识才能对其加以区分。计算机无法评估事物是真正相关还是偶然相关，因为计算机不理解数据的意义。数字不过是数字而已。计算机并没有区分好坏数据所需的人类判断力，没有分辨有理有据的和虚假伪造的统计学模型所需的人类智能。如今的计算机能通过图灵测试，却无法

通过史密斯测试。如果所发现的模式被隐藏在黑匣子里，让模型难以理解，就会使这种情况恶化。无人知晓为什么计算机算法决定要买入这只股票、拒绝这名求职者、给病患开这种药、拒绝这名囚犯的假释请求和轰炸这座建筑。

在大数据时代，真正的危险不是计算机比我们更聪明，而是我们自己这么认为，从而信任计算机为我们做出重要决定。我们不应该认为计算机就是万无一失的、数据挖掘都是“知识发现”，以及黑匣子也应该被信赖。我们要相信自己能判断统计学模型是否合理，有无可能派上用场，抑或它只是偶然出现而已，转瞬即逝，毫无用处。

人类推理与人工智能有天壤之别，这也是为什么如今更显人类推理能力的可贵之处。

